cu6-H05946-10-kp1115

THE JOURNAL OF
PHILOSOPHY

10

VOLUME XCVIII, NUMBER 1
JANUARY–DECEMBER 2001

Edited by
Brian Barry, Bernard Berofsky
Akeel Bilgrami, Arthur C. Danto
Kent Greenawalt, Patricia Kitcher
Philip Kitcher, Isaac Levi
Mary Mothersill, Philip Pettit
Carol Rovane

Consulting Editors
David Albert, John Collins
James T Higginbotham, Charles D. Parsons
Achille Varzi

Managing Editor
Michael Kelly

Published by The Journal of Philosophy, Inc.

CALCUTTA UNIVERSITY LIBRARY
Acc. K.P.-1115
Date 19.09.02
ALIPORE CAMPUS

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us *Each copy must include a notice of copyright* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying Inquiries about photocopying may be sent to the Managing Editor

© Copyright 2001 by the Journal of Philosophy, Inc ISSN 0022-362X

POSTMASTER Second class postage paid at New York, NY, and other mailing offices Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, NY, NY 10027 Fax (212) 932-3721

THE JOURNAL OF PHILOSOPHY

5/2001

VOLUME XCVIII, NUMBER 1
JANUARY 2001

page
5 *Knowing Numbers* Marcus Giaquinto
19 *The Rationality of Science and the Rationality of Faith* Theodore J Everett

BOOK REVIEWS
43 Elliott Sober and David Sloan Wilson *Unto Others The Evolution and Psychology of Unselfish Behavior* P Kyle Stanford
47 Carol Rovane *The Bounds of Agency An Essay in Revisionary Metaphysics* Catherine Wilson

VOL|P-37

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022–362X

POSTMASTER. Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

04 5/2001

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

VOLUME XCVIII, NUMBER 2
FEBRUARY 2001

page
55 *The Principle of Sufficient Reason* Gordon Belot
75 *Causes as Probability Raisers of Processes* Jonathan Schaffer

BOOK REVIEWS
93 John Martin Fischer and Mark Ravizza *Responsibility and Control A Theory of Moral Responsibility* Michael McKenna

101 MEMO TO AUTHORS
102 NEW BOOKS

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc ISSN 0022–362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS
5|5|2001

VOLUME XCVIII, NUMBER 3
MARCH 2001

page

111 *Disjunctive Properties Multiple Realizations* Lenny Clapp
137 *The End of Millianism Multiple Bearers, Improper Names, and Compositional Meaning* Jerrold J Katz

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022-362X

POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

25/5/2001

VOLUME XCVIII, NUMBER 4
APRIL 2001

page
167 *Systematicity and the Cognition of Structured Domains*
Robert Cummins, James Blackmon, David Byrd,
Pierre Poirier, Martin Roth, Georg Schwarz
186 *Personal Identity, Memory, and Circularity An Alternative for Q-Memory* Marc Slors

215 NEW BOOKS ANTHOLOGIES
218 NEW BOOKS TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022-362X

POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

28\6|2001

VOLUME XCVIII, NUMBER 5
MAY 2001

page
219 *The Most Famous Equation* Marc Lange
239 *Does Perception Have a Nonconceptual Content?* Christopher Peacocke

BOOK REVIEWS
265 Noam Chomsky *New Horizons in the Study of Language and Mind* Gilbert Harman
269 John Rawls *Collected Papers* Thomas E Hill, Jr

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

18\7\2001

VOLUME XCVIII, NUMBER 6
JUNE 2001

page
273 *The Intransitivity of Causation Revealed in Equations and Graphs*
Christopher Hitchcock
300 *Paradoxes of Fair Division*
Steven J Brams, Paul H Edelman, and Peter C Fishburn

BOOK REVIEWS
315 Jaegwon Kim *Mind in a Physical World An Essay on the Mind-body Problem and Mental Causation* Barry Loewer
325 Peter Ludlow *Semantics, Tense, and Time* Ned Markosian

330 NOTES AND NEWS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022-362X

POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J E Woodbridge and J McKeen Cattell In 1906, Wendell T Bush became associated with the JOURNAL as co-editor The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923 Since its founding, it has been published from the Department of Philosophy at Columbia University It is typeset and printed by Cadmus Journal Services

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO. 1, JANUARY 2001

64 5/2001

KNOWING NUMBERS

Simple arithmetic entails that there are numbers, two of them between 3 and 6, for example, and infinitely many thereafter. But how do we know that there are any such things? If numbers exist, they would surely be abstract, and there seems to be no way of explaining our knowledge of abstract objects, short of postulating some supernatural mode of apprehending them. Numbers in particular do not emit or reflect signals, they leave no traces, their behavior causes no phenomena from which their existence may be inferred. Although it is plausible that arithmetic is indispensable to empirical science, this has been disputed, and even if true, its significance for the truth of arithmetic and the existence of numbers is moot, as one might take an instrumentalist view of arithmetic in science.[1] Considerations of this sort have led some people to conclude that we do not have knowledge of numbers, and to seek to accommodate the prevalence and utility of arithmetic in one way or another within a nominalist framework.

---

[1] Hartry Field has argued against the claim that arithmetic is indispensable to science—*Science without Numbers* (Princeton University Press, 1980). Some accept the claim but dispute its epistemological significance: Penelope Maddy, "Indispensability and Practice," this JOURNAL, LXXXIX, 6 (June 1992): 275-89, and Elliott Sober, "Mathematics and Indispensability," *Philosophical Review*, CII (1993): 35-57. Michael Resnik defends both the claim and its significance—*Mathematics as a Science of Patterns* (New York: Oxford, 1997). The indispensability criterion is stated in Hilary Putnam, "Mathematics without Foundations," in *Mathematics, Matter, and Method: Philosophical Papers*, Volume I (New York: Cambridge, 1967). Putnam attributes the criterion to W. V. Quine.

0022-362X/01/9801/5–18

Nominalism is the view that there are no abstract entities[2] Nominalism about arithmetic is the view that either numbers are not abstract, or arithmetical sentences, correctly interpreted, do not entail the existence of numbers, or arithmetic is false Of these three brands of nominalism, the first is not popular[3] and so I shall say no more about it The task for the second brand is to find a plausible interpretation of arithmetic that maintains truth and avoids commitment to abstract entities[4] A route favored recently is to go modal interpret arithmetical statements as about possibilities or possible concrete objects A common feature of these attempts is that under the chosen interpretation, the proposition expressed by an arithmetical sentence differs from anything we have in mind when using the sentence[5] So this brand of nominalism is prima facie implausible, and it leaves open the question of the truth of the propositions we actually have in mind when using arithmetical sentences

The third brand of nominalism about arithmetic is fictionalism According to fictionalism, arithmetical sentences say just what we normally think they say, and our arithmetical theorems have the existential consequences we normally think they have, but arithmetical theorems with existential consequences are one and all false[6] There are infinitely many primes? False! 2 + 1 = 3? False! This is uncomfortable Simple numerical equations are among the propositions that seem most obviously true There are various emollients to alleviate discomfort arithmetic does not have to be true to be useful, and we can specify a kind of correctness without appeal to truth derivability from axioms Moreover, there are various reconstructive strategies[7] for finding a nominalistic statement capturing "the grain of truth" in a theorem of arithmetic But this satisfies only a lucky few Surely, some misunderstanding is involved in disbelieving that $2 + 1 = 3$, for example? Thus we may be trapped, stuck at one or another unsatisfactory position or oscillating between them

---

[2] What this comes to is not clear, as the abstract-concrete distinction is hazy But it is widely assumed that numbers would have to be abstract For the difficulty of drawing the abstract-concrete distinction, see Bob Hale, *Abstract Objects* (Cambridge Blackwell, 1987), chapter 3

[3] The only one I know of who might qualify as a nominalist of this kind is Bishop Berkeley, who claimed that "numbers are nothing but names"—*Philosophical Commentaries* (New York Garland, 1989), § 763, cf § 803 It is not clear what he would have said about numbers beyond those for which a name has been uttered or written down

[4] This is dubbed "the hermeneutic strategy" in John P Burgess and Gideon Rosen, *A Subject with No Object* (New York Oxford, 1997)

[5] For example, on one nominalist interpretation of arithmetic, an arithmetical sentence *S* is to be understood roughly as follows there could be a model of the second-order Dedekind-Peano axioms for arithmetic and necessarily, for any model *M* of those axioms, *S* is true in *M* Clearly, only someone with some knowledge of mathematical logic would even entertain a thought of this kind For details, see Geoffrey Hellman, *Mathematics without Numbers Towards a Modal-Structural Interpretation* (New York Oxford, 1989) For an alternative nominalist view, see Charles Chihara, *Constructibility and Mathematical Existence* (New York Oxford, 1990)

[6] See Field, and David Papineau, "Knowledge of Mathematical Objects," in A D Irvine, ed , *Physicalism in Mathematics* (Boston Kluwer, 1989), pp 154-81

The aim here is to show that this trap is avoidable, by arguing for the possibility of a naturalistic account of knowing numbers, without taking them to be concrete objects I shall try to show that cognitive science already provides some of the resources for such an account The numbers I have in mind are the finite cardinals some we know by acquaintance and some by description The paper takes these cases in order, and then something is said about knowing finite numbers more abstract than cardinals But first I want to say what cardinal numbers are, and to present a general setting for the epistemology to follow

I WHAT ARE CARDINAL NUMBERS?

Cardinal numbers are properties In my view, they are properties of sets, but they might also be properties of concept extensions, collections, pluralities, nonmereological aggregates, or some other kind of collective, provided that collectives of one or zero items are not excluded I shall talk of sets for convenience, but I do not insist that only sets have cardinal number As long as one is clear about which are the items in the collective to be numbered, one need not deploy a general collective concept in framing the question What is the number of sheep in this flock? In fact, one can do without any collective concept, general or specific How many sheep are here? Nonetheless, the answer to the question gives the cardinal size of the flock, and that is as much a property of the flock as its monetary value or the average age of its members

But do any properties exist? Are properties and kinds not just a philosopher's fiction, or the result of a mistakenly literal reading of certain *façons de parler*? This view was famously expressed by W V Quine[8]

> One may admit that there are red houses, roses and sunsets, but deny, except as a popular and misleading way of speaking, that they have anything in common [T]he word 'red' or 'red object' is true of each of sundry individual entities which are red houses, red roses, red sunsets,

---

[7] Burgess and Rosen survey the reconstructive strategies

[8] "On What There Is," in his *From A Logical Point of View* (Cambridge Harvard, 1953), pp 1-19

> but there is not, in addition, any entity whatever, individual or otherwise, which is named by the word 'redness', nor, for that matter, by the word 'househood', 'rosehood', 'sunsethood' (*ibid.*, p. 10).

Plato[9] may have thought that for every general term there is a corresponding Form. Quine held that for no general term is there a corresponding property or kind. Between these two extremes lies the "mixed" view that for some but not all general terms does there exist a corresponding property or kind. Although there is no decisive argument for any of these views, I hold that the mixed view should be our default hypothesis. It is true that we rarely assert or deny the existence of a property or kind, as opposed to instances of a property or kind. But although talk of this kind is rare, judgments about the existence or nonexistence of properties and kinds are sometimes made in science. For example, Joseph Priestley thought that all combustible material contained phlogiston, a substance that is liberated from the material in combustion, with the dephlogisticated substance left as an ash or residue. Antoine Lavoisier thought that there was no such substance as phlogiston and no such property as being dephlogisticated.

Is this really a case of a difference about the existence of a kind or property, rather than about the existence of instances? Of course, they did disagree about the existence of instances. But there was more to it than that. On Lavoisier's view, nothing *could* contain phlogiston or be dephlogisticated. Compare this with the case of a transuranium element prior to evidence that there exist any atoms of that element. Before 1996, no sample of the element with atomic number 112 had been found and, as far as was known, there was none in existence. Yet it was held that there could be some, and in fact some was produced (by bombarding lead with zinc). This is the current situation for the element with atomic number 114.[10] Thus, scientists view the following as an epistemic possibility: there is no sample of element 114 but some could be produced. Although it is also believed that there is no sample of phlogiston, it is not also believed that some could be produced. What explains the difference? Surely, it is not that present technology for producing phlogiston lags behind present technology for producing superheavy elements, nor even that there is an insuperable technological barrier, such as an unattainable energy requirement, to phlogiston production The answer is that while there is such a kind as element with atomic number 114 and such a property as having a nucleus with 114 protons, there is no such kind as phlogiston and no such property as being phlogiston

---

[9] *The Republic*, G. M. A. Grube, trans. (Indianapolis: Hackett, 1992) 596a. *X corresponds to* general term *T* when *X* is designated by the singular term formed from *T* (for example, *T*-ness, or in Plato's writing, the *T* itself).

[10] By the time you read this, the claim may be out of date, but the philosophical point is unaffected.

Are we being misled by popular ways of speaking here? Presumably, we have a naive linguistic disposition to generalize existentially on a noun phrase, and the idea that kinds and properties exist might be nothing but an unconscious conversion of this disposition into an ontological thesis, regardless of empirical justification But in holding the mixed view (that some noun phrases for properties or kinds do designate while others do not), we clearly resist that disposition On the mixed view, decisions about what kinds and properties are real is to be based on empirical findings The extreme views, by contrast, issue from metaphysical prejudice, as they are unlikely to be accepted by anyone without a platonist or a nominalist bent[11] For these reasons, I am going to take the mixed view as the default position and use it as a background premise That of course leaves open the question whether there are such properties as cardinal numbers The affirmative view is not confined to those with a general predisposition to platonism John Locke,[12] for example, included number in the list of what he was prepared to call "real qualities " But this view needs vindicating The challenge is to show how, if there were cardinal number properties, we could have knowledge of them I shall try to meet that challenge in what follows by sketching a naturalistic account of such knowledge

II ACQUAINTANCE WITH CARDINAL NUMBERS

We say that we know John Doe, but for entities other than people, talk of knowing them is unusual and sometimes odd But not always on hearing the opening bars of a symphony, one might say 'I know that music', on seeing the handwritten address on an envelope, one might say 'I know that handwriting' Here one claims to recognize the music or handwriting, and if the claim is true one must be acquainted with it[13] In these cases, acquaintance entails that one has already heard performances or seen samples, and from this prior experience one has acquired a concept of the music or handwriting, and an ability to recognize the music or handwriting in other performances or samples, where this involves thinking of them under a relevant concept Generalizing, I shall take it as sufficient for acquaintance with a property that one has experienced instances of it, and as a result one has acquired a concept of the property and an ability to recognize other instances and to discriminate them from noninstances (unless circumstances are unfavorable), where this involves applying the concept exclusively to the instances[14]

---

[11] At the time Quine wrote "On What There Is," he was not the (reluctant) Platonist that he later became That article was first published in 1948, a year after the publication of "Steps towards a Constructive Nominalism," *Journal of Symbolic Logic*, XII (1947) 105-22, co-authored with Nelson Goodman In that article they declare their belief in the nonexistence of abstracta to be a "philosophical intuition that cannot be justified by appeal to anything more fundamental" (p 105)

[12] *An Essay Concerning Human Understanding* [1689] (New York Oxford, 1975), Book II, Chapter VIII, § 17 One might argue that what Locke really had in mind here is not the number *n* but the property of having *n* atomic constituents ("solid parts") But this reading of the passage is dubious and cannot easily be transferred to other passages in which Locke seems to say that number is a primary quality

What does recognitional ability consist in when the property concerned is a cardinal number? It is not enough to be able to tell whether a given set has the cardinality By careful counting, we can tell whether a given set of stationary visible objects whose cardinality lies somewhere in the 200-to-300 range has cardinality 271 That would be mere detection rather than recognition Recognition of *n* requires that we have some sense of the cardinal size *n* as distinct from its neighbors For which cardinals do we have such a sense? There is evidence that we have an innately given magnitude representation of rough cardinal size, or *numerosity*, with a neural basis in the inferior parietal lobes[15] The evidence comes from a variety of sources experiments on healthy adults and children, clinical tests on brain-damaged patients, brain-imaging techniques, and studies on animals from parrots to primates[16] The capacity for representing numerosities shares a couple of features with other innate rough-quantity senses, such as sense of duration, temperature, and pitch These features are known as the *distance effect* and the *magnitude effect* Crudely put, the distance effect is that the smaller the difference between two quantities (for a fixed mean), the harder it is to distinguish them, the magnitude effect is that the greater the mean of two quantities (for a fixed difference) the harder it is to distinguish them[17] Difficulty of discrimination is measured by response times and error rates for tasks requiring comparison of the quantities The pattern of response times and error rates is part of the data to be explained, the fact that this pattern is the signature of rough-quality sensing supports the view that in making number judgments of the kinds, tested subjects are sensing rough qualities[18] Putting together the distance effect and the magnitude effect, it follows that in using our sense of numerosity to distinguish adjacent numbers, it will be easier for small numbers, and this is how it seems to be There is evidence from studies with very young children and animals for a prelinguistic ability to discriminate cardinal numbers 1, 2, and 3 This ability may be provided by our numerosity sense, something that is predicted by a neural network model for this sense[19] This prelinguistic ability is matched by an extremely fast and reliable ability in adults to sense the cardinal numbers of sets of 1, 2, and 3 visually presented items, known as *subitizing*[20] So we can sense these cardinal numbers, and once we have concepts for these cardinal numbers, we can recognize instances of them and discriminate them from noninstances (in particular, from instances of adjacent cardinal numbers), as we do in subitizing.

---

[13] I take it as obvious that one would not be claiming to know that very performance of the symphony or those very ink marks on the envelope What one claims to know is the type rather than any of its tokens

[14] An anonymous referee points out that this account of property acquaintance sounds perfectly acceptable to an antirealist about properties who favors conceptualism While this may be true, it is not in conflict with my aim at this point, which is to argue that, if there were cardinal-number properties, we could have knowledge of them My reason for taking seriously talk of properties is given in the previous section

[15] This is what Stanislas Dehaene refers to by the title phrase of his book *The Number Sense* (New York Oxford, 1997)

[16] See Brian Butterworth, *The Mathematical Brain* (London Macmillan, 1999), chapters 3-6 This was published in the United States as *What Counts How Every Brain Is Hardwired for Math* (New York Free Press, 1999) See also Dehaene, chapters 1-3, 7, 8

---

[17] For precise statements of these features, see any textbook in psychophysics under "Fechner's Law" and "Weber's Law "

[18] Thus it is a mistake to think that the only data we have are data to the effect that we know facts of the form 'There are *n* so-and-sos' for small *n* This, I hope, partly answers the question of an anonymous referee how such data justify the claim that we know numbers as objects In fact, I do not want to claim that we know numbers as objects, and if some formulations suggest otherwise, please regard that as a slip My claim is that a naturalistic account of knowledge of numbers is not impossible if we take cardinal numbers to be properties

[19] Dehaene and Jean-Pierre Changeux, "Development of Elementary Numerical Abilities A Neuronal Model," *Journal of Cognitive Neuroscience*, v (1993) 390-407 The hypothesis that the numerical abilities of animals and infants are based on a sense of numerosity is put forward by Dehaene in *The Number Sense* among others This is disputed in Susan Carey, "Evolutionary and Ontogenetic Foundations of Arithmetic," *Mind and Language* (forthcoming) Carey's view is that we have the sense of numerosity but that performance of numerical tasks by animals and infants uses other cognitive resources

[20] Subjects are asked to indicate the number of presented items flashed on a screen as quickly as they can When response time is plotted as a function of number of items viewed, there is a discontinuity in slope almost horizontal for small numbers (too fast for silent counting), steeper for larger numbers (matching increments in counting time) Error rates show the same pattern The discontinuity persists under diverse display conditions This is evidence for different processes of enumeration a fast, error-free process for small numbers, that is, subitizing, and a slower, less reliable process for larger numbers (The subitizing limit varies between 3 and 7, depending on type of experiment, visual display, statistical techniques used to calculate subitizing limit, and individual differences)

What kind of experiences would help us get concepts of cardinal numbers?[21] For small nonzero cardinal numbers, our experiences of counting would be a natural first call. But one has to be careful here. What does counting consist in? On some definitions, counting might prerequire concepts of cardinal numbers, in which case our acquisition of the concepts could not be the result of counting experience. Let us take counting to include the one-to-one assignment of words in a stable order with a perceptually given set of entities. That is, to be a competent counter a child must obey the stable-order principle and the one-one principle (allowing for slips): the counting words must be used in the same order in different counts and they must be assigned one-to-one with the items being enumerated.[22] A child who obeys the stable-order and one-one principles may still not know what counting is for. Even when a child knows that the correct answer to the question 'How many?' is given by the last word in the count, she may not understand the question as a request for the number of items counted.[23] Experimental results show that children come to realize that the last word in a correct count signifies the number of the objects counted only after they have acquired the stable-order principle and the one-one principle. That is, they learn the cardinal significance of counting only after learning correct counting procedure.[24] At about the same time, they come to know which cardinal number each word in their counting list denotes.[25] I assume that in the process they acquire concepts for those cardinal numbers. How this happens is not clear. It is not improbable that part of the process involves mentally associating representations of initial words in the counting list with initial representations supplied by the sense of numerosity.[26] An additional possibility is that in learning to count we construct a category representation of sets of a given size, one for each set size from 1 to 3.[27] These might then serve as representations of those cardinal numbers and get mapped onto the initial numerosity representations. Our counting experience might also lead us to associate adding a further item to the set of those counted with proceeding from one cardinal number to the next one up in order of size. So our concept of cardinal 2, for example, would present it as the successor of 1 as well as the number between 1 and 3. Thus, experience of the counting procedure and of its use may feed into the acquisition of concepts for cardinal numbers 1, 2, and 3, and enable us to recognize new instances of each and discriminate them from instances of the neighboring cardinal numbers. In this way, we can have acquaintance with these numbers.

---

[21] It is possible that we have innate concepts of cardinals 1, 2, and 3. Experiments of Karen Wynn suggest that infants as young as five months make numerical predictions when numbers are limited to three—"Addition and Subtraction by Human Infants," *Nature*, CCCLVIII (1992): 749-50. Wynn's findings have been replicated many times under varying conditions. On one explanation of these findings, infants have genuine numerical beliefs, which have concepts for the cardinals 1, 2, and 3 as components. The explanation of these findings is under dispute, however.

[22] For counting principles, see Rochel Gelman and Charles Gallistel, *The Child's Understanding of Number* (Cambridge: Harvard, 1978), chapter 7.

[23] Children at this stage sometimes treat a number word as a variable designator, like a demonstrative, taking it to refer in a given count to the item to which it is assigned. See Karen Fuson, *Children's Counting and Concepts of Number* (New York: Springer, 1988), and Wynn, "Children's Understanding of Counting," *Cognition*, XXXVI (1990): 155-93.

[24] Wynn (*ibid.*) found that the children in her study acquire the cardinal principle at around the age of 3 ½ years, after they have learned correct counting procedure. This is contrary to the understanding-before-skill hypothesis of Gelman and Gallistel, but accords with the reverse hypothesis of Fuson.

[25] Wynn, "Children's Acquisition of the Number Words and the Counting System," *Cognitive Psychology*, XXIV (1992): 220-51. Wynn found that children take the

Do we have acquaintance with cardinal numbers greater than 3? My inclination is to think that we do. In our experience of counting, we repeatedly meet and notice numbers from 4 to 10, as finger counting typically has an important developmental role.[28] In so doing, we might sharpen our sense of numerosity to get senses of individual cardinal numbers beyond 3, just as repeatedly exercising

---

number words in their list to stand for unique cardinalities before knowing which cardinality each one denotes. They may reach this stage through experience of the use of number words outside counting contexts, reaching the later stage only when they make an order-preserving association of number words in counting sequence with cardinals in order of size, perhaps by some general analogizing mechanism. See also Fuson, "Relationships Between Counting and Cardinality from Age 2 to Age 8" in Jacqueline Bideaud, Claire Meljac, and Jean-Paul Fischer, eds., *Pathways to Number* (Hillsdale, N.J.: Lawrence Erlbaum, 1992), pp. 127-49.

[26] There is clinical evidence for neural links between the sites in the left hemisphere of the numerosity sense and of the aural number word store (inferior parietal lobe and perisylvian cortex). See Butterworth, chapter 4, and Dehaene, *The Number Sense*, chapter 7. Presumably these links are formed when children learn the cardinal meanings of the number words.

[27] Perhaps by "abstracting" from the nature of the elements of *n*-membered sets and the order in which the members are given, as Georg Cantor suggested, we form a representation of the cardinal number *n*—*Contributions to the Founding of the Theory of Transfinite Numbers* I (1895), Philip E. B. Jourdain, trans. (New York: Dover 1955), §1. This is not to endorse Cantor's view that these representations are the cardinal numbers.

[28] See Fuson, "Relationships between Counting and Cardinality from Age 2 to Age 8." For the neuropsychological importance of finger counting, see also Butterworth, chapter 5.

one's visual capacity for discriminating shades of red or types of snow sharpens our sense of those different properties or kinds From counting experience, we get a sharpened sense of the cardinal size 4 Our sense of 4 may not be quite as clear and strong as our sense of 3, but the difference may not be great Similarly, we may develop a sharpened sense of subsequent numbers, each almost as clear and strong as our sense of its predecessor It is possible that there is a relatively sudden drop in the sharpness of the sense of numerosity thus developed, perhaps after 10, but I am not aware of evidence for it If the contrary is true, there would be no natural cut-off point separating those cardinals we can recognize and those we can only detect, consequently, acquaintance with cardinals would have to be a matter of degree But there is no ground for worry here, as that is how it is anyway for acquaintance with material individuals

III KNOWLEDGE OF CARDINAL NUMBERS BY DESCRIPTION

Small numbers beyond 3 may be known by acquaintance But even if they are not, they are known by description through familiarity with small number addition The cardinal number 5, for example, may be known as the size of a set composed of a triple and a pair (without repetition or overlap) Clearly, all the numbers from 4 to 10 can be easily and briefly described this way without assuming knowledge of cardinal numbers beyond 3 Early addition exercises performed using counting techniques give us knowledge of all two-addend cardinal sums with totals up to and including 10 [29] Thus, for each number from 4 up to 10 there are several descriptions of this sort (sum of cardinal *x* and cardinal *y*) by which we know it, and we have more of these descriptions for larger numbers, so that our richer descriptive knowledge of larger numbers makes up for our weaker direct sense of them [30]

This kind of knowledge takes us up to 10, perhaps to 20, perhaps even to 100 What about knowledge of much larger numbers? There are many possibilities, but probably just a few basic kinds of description that most of us use [31] As multidigit numerals (in the decimal place system) play such a big role in numerical thinking, it is natural to look to them for clues An obvious thought is that a multidigit numeral is short for the corresponding polynomial description the cardinal 271 is the cardinal of a set composed of 2 hundreds, 7 tens, and 1 one This seems plausible and can clearly be extended using greater powers of ten But we would need to say how we know of those powers of ten This is easy for the case of a hundred, which is thought of as a set of 10 tens But for higher powers the question arises whether our knowledge of the cardinal number really depends on its description in terms of powers Do we think of the cardinal 2,710,003 as the cardinal of a set composed of 2 tens-to-the-power-of-six, 7 tens-to-the-power-of-five, and so on? Do we not first think of a ten-to-the-power-of-six as the number of things denoted by a '1' followed by six '0's? The actual story may be quite complicated But whatever it is, we clearly *could* come to know cardinal numbers beyond 10 under the polynomial description, using a recursive definition for powers of ten a ten-to-the-power-of-two is 10 tens, a ten-to-the-power-of-n+1 is 10 tens-to-the-power-of-n

---

[29] For a developmental account, see Fuson, "Relationships between Counting and Cardinality from Age 2 to Age 8 "

[30] The cardinal number 0 may (must?) be known by description—for example, the number of wild tigers in Scotland, the result of taking away two things from a set of two Perhaps one does not think of '0' as naming a cardinal number (rather than an indicator of absence) ahead of accepting that there is an empty set I do not know of empirical work on this question

[31] I have in mind people who are not calculating prodigies but have a level of schooling in arithmetic that is standard in North America and Europe I am not suggesting that this standard is very good or better than standards elsewhere

On the suggestion just made, we know larger numbers under their descriptions as polynomials, with multidigit numerals serving as the descriptive expressions There is an alternative, in which the multidigit numerals play a more intimate part in our knowledge of numbers the numerals are not merely the carriers of the description but are part of the descriptive content. How is this going to avoid circularity? If we think of the number 27176 as the number denoted by '27176', that cannot be the whole story Through our knowledge of an algorithm for determining the numeral immediately preceding any given multidigit numeral and our possession of a concept for the cardinal successor function, we can amplify as follows the number denoted by '27176' is the cardinal successor of the number denoted by the numeral predecessor of '27176' Of course, this works only for someone who has knowledge of the number denoted by '27175' But the same kind of description is available, using the following general schema the number denoted by a multidigit numeral μ is the cardinal successor of the number denoted by the numeral predecessor of μ Repeated application brings us back to single-digit numerals whose denotation we know by direct association with cardinals with which we are acquainted In this way, circularity is replaced by recursiveness I am not suggesting that we actually run back in thought from a multidigit numeral '27176' to single digits in thinking of the number it denotes We just think of the number as the number denoted by '27176', but our understanding of this description (in terms of numeral predecessor and cardinal successor functions) suffices to fix its reference That is an alternative possibility[32]

Probably the actual way in which we think of multidigit numbers is more complicated than either of the two suggestions made here, perhaps involving a mixture the polynomial description for numbers of two to four digits and the apparently circular description in terms of numerals for larger numbers Even this is liable to be an oversimplification, as it ignores how natural-language expressions for larger numbers affect how we think of them On top of this there will be variations among people and over time But my aim here is not to give "the correct account" of our knowledge of numbers by description It is merely to show that such an account is possible without positing powers that could not be countenanced in cognitive science

IV KNOWING THE OBJECTS OF PURE NUMBER THEORY

Even if my case for the possibility of a naturalistic account of knowing cardinals is accepted, one might protest that we still lack a parallel case for the objects of number theory Those objects are not set sizes, they are, if anything, positions in the natural number structure[33] I am sympathetic to this view Pure mathematicians at work in number theory are not concerned with finite cardinals as opposed to finite ordinals and the myriad other systems that exemplify this structure, they are concerned with the structure itself Hence numerals and variables in the context of pure number theory are used to refer to (or range over) positions in the structure But how could anyone have knowledge of such highly abstract objects, positions in a structure?

Although it lies beyond the scope of this paper to answer this question in any detail, I think an outline is possible First, we would have to have knowledge of the structure itself Assuming that we cannot have acquaintance with more than a finite portion of an infinite structure, knowledge of the structure must be knowledge by description, a description whose components are inferred from other knowledge One possibility is that we derive a description of the natural-number structure from our knowledge of the system of numerals in order of precedence, which is one of the systems having that structure I assume that we are acquainted with some numerals (types, not tokens) and that we know about the system as a whole through our knowledge of algorithms for constructing the successor and predecessor of a given numeral, and for addition on the numerals Writing $|\kappa|$ for the number of predecessors of numeral κ,[34] $\mu + \kappa$ is the numeral constructed by $|\kappa|$ applications of the successor operation starting from numeral μ Precedence is understood in the familiar way $\mu < \nu$ if and only if for some κ other than '0', $\mu + \kappa = \nu$ Through standard arguments we know that the order of precedence is a strict total ordering having just one element that is not a successor and no element that is not a predecessor In fact, it is a well ordering, because from any numeral μ one can reach '0' by $|\mu|$ applications of the predecessor operation, and so there is no unending sequence of ever earlier numerals, hence every nonempty set of numerals has an earliest member With this knowledge of the structure, as a well-ordering with one initial[35] position and no terminal position, we can know a position as the n^{th} position of the structure For any numeral not too long to be grasped, we can know and name the position that it occupies and all preceding positions

---

[32] Neither of these accounts is intended to show how we might know all finite cardinals That would be impossible, as there is a finite upper bound on the number of possible neural states, hence on the number of representations (sensory or descriptive) that the mind can grasp

[33] More than one structure may be described as "the" natural-number structure, depending on the relations considered For simplicity, we can take the sole relation to be the less-than relation For structuralist views of the ontology of pure mathematics, see Stewart Shapiro, *Philosophy of Mathematics* (New York Oxford, 1997), part II, and Resnik, part III

That is my outline I concede that we are still some way off from being able to fill in the cognitive details of a naturalistic account of our knowledge of the natural-number structure, hence of its positions But I am aware of no reason to think that future developments in the empirical study of our cognitive capacities will not one day provide the resources for a full account along the lines of the sketch just given

V CONCLUSION

Among philosophical attitudes to arithmetic prevalent today is a kind of skepticism the theorems of arithmetic are not true, unless they are interpreted in one or another strange way One of the main motivations for this attitude is the conviction that there could not be a naturalistic mode of knowing numbers I have tried here to rebut that conviction Drawing on recent work in cognitive science, I have sketched ways in which we might have knowledge of cardinal numbers I have not canvassed all the possibilities, and it may be that none of those presented here are right My claim is only that they, or something similar, could be right, so, I conclude, we should not dismiss the very possibility of explaining knowledge of cardinal numbers within a scientific framework.

---

[34] I assume that '0' is the first numeral So $|\kappa|$ is the cardinal denotation of κ

[35] This terminology is slightly misleading An *initial* position is one that is not the successor of another position So the position of any limit ordinal is initial (as is the position of 0)

What has prevented people from seeing the kind of possibilities outlined here? There are probably several obstacles. Here are my suggestions. One has to do with the word 'object'. If cardinal numbers exist, they are objects, we say. That dictum is a truism if the word 'object' is used to mean 'entity'. But cardinal numbers, being properties of sets, are very unlike prototypical objects, which are bounded, perceptible, spatiotemporal continuants. Moreover, the word 'object' is often used contrastively with 'property'. So the dictum may obscure the fact that cardinal numbers are properties and incline one to think that, if there were such things as cardinal numbers, they would have to be something like disembodied billiard balls. A second obstacle may be a nominalist reluctance to accept the existence of properties. A third obstacle may be that the empirical study of cognitive capacities for number knowledge is relatively young—though it is now a vibrant research subject. Finally, philosophers of mathematics have shown a near total disregard for the empirical findings, despite a willingness to pronounce on cognitive matters.[36] I hope that this paper shows the merit of abandoning that attitude.

MARCUS GIAQUINTO

University College London

---

[36] A notable case is Philip Kitcher's claim that children come to learn the meaning of 'number' and to accept basic truths of arithmetic "by engaging in *activities* of collecting and segregating"—*The Nature of Mathematical Knowledge* (New York: Oxford, 1984), chapter 6. Kitcher gives no reference to empirical work on children's cognitive development, and his view is not borne out by the findings earlier published in Gelman and Gallistel.

THE RATIONALITY OF SCIENCE AND THE RATIONALITY OF FAITH*

If science is rational, why is it so rare in human history? If faith is *irrational*, why have the great majority of people always preferred traditional beliefs to new ones obtained by independent, critical investigation? The prevailing answer is that nonscientific beliefs derive primarily from nonrational causes, and, unfortunately, the history of the world is largely dominated by unreason. I think that this answer is false. I see the conflict between science and faith as resulting not from the struggle of rational thought against irrational forces of conformity, but from a systematic tension between two aspects of rationality itself, which I call simply subjective and objective. I make two main controversial claims. I claim that most people rationally (in the more basic, subjective sense) *ought* to believe in their local traditions, because that is what all of the people around them believe, and few individuals are in a total epistemic situation from within which they can reasonably contradict their neighbors. I also claim that most scientists and other modern intellectuals ought rationally (in the same sense) *not* to believe in their own theories, though it is often good for others in the long run when they do. This is not to deny that scientific research is objectively a rational process, that is, one that leads to true beliefs for people in general. But I claim that this requires a kind of epistemic altruism on the part of experimental thinkers, in that their new ideas are ordinarily less likely (from the thinkers' own, internal points of view) to be correct than are the opinions they have received from others.

My argument has the following structure. In section I, I specify the difference between my conceptions of subjective and objective rationality, and make a few other preliminary distinctions that are needed for a clear exposition of my main points. In section II, I defend as subjectively rational a broad principle of deference to the beliefs of others, including most traditional religious beliefs. I also consider two main objections, one derived from David Hume and one from René Descartes. In section III, I discuss the combination of objective rationality and subjective irrationality which accounts both for the

---

* I would like to thank Larry Blackman, Ellery Eells, Berent Enç, Richard Fumerton, Alan Sidelle, Walter Soffer, Dennis Stampe, and Stephen Sullivan for helpful comments on this paper. I would also like to thank SUNY-Geneseo for its support through a Presidential Summer Fellowship.

0022-362X/01/9801/19–42

great success of science, and for its being so rare. In section IV, I consider briefly whether this new view leads into paradox when applied to itself.

I. DISTINCTIONS

There are three intuitive distinctions which I want to discuss in advance of my main arguments: subjective versus objective rational belief, first- versus second-order rational belief, and rational belief versus knowledge. There is nothing very new or very technical here; I just want to avoid certain natural misunderstandings.

Pure rationality itself is a single, formal thing—a connection between bodies of evidence and sets of beliefs, through chains of proper inference. But a useful distinction can be drawn between different applications of this single concept, in terms of the available bodies of evidence. I say that a belief *p* is *subjectively rational* for a person *S* at a time *t*, if *p* is what *S* rationally ought to believe at *t*, given all of the evidence available to *S* at or before *t*. I do not mean whatever *seems* to the person to be rational. I mean what someone really ought to believe, not just what he thinks that he ought to believe, given all of the evidence that he has had so far. This strong conception will rule out as fully rational some people who have made errors in reasoning in the past that they cannot now correct. Weaker notions of subjective rationality are no doubt worth exploring, but one of my goals here is to make a strong defense of traditional believers, so I am requiring them to be completely "epistemically responsible" agents, beyond any relevant criticism respecting the rationality of their background beliefs.[1]

I shall not try to define 'rationally ought to believe' more precisely; common intuitions should be good enough.[2] My immediate point is just that different people frequently have access to different pools of evidence, so it can be equally rational for them to come up with different beliefs All such beliefs, if they are properly inferred from all the evidence available, will be subjectively rational for the people who hold them

---

[1] Some writers have used the terms 'objective' and 'subjective' to mark the difference between *really* and *apparently* rational belief, that is, what merely seems to someone to be rational belief. Richard Feldman has argued plausibly that this distinction tends to collapse when looked at closely, unlike a similar distinction in ethics between really and apparently right action—"Subjective and Objective Justification in Ethics and Epistemology," *Monist*, LXXI (1988): 405-19.

[2] A quick guess would be that *S* ought rationally to believe *p* just in case *p* has a subjective probability for *S* of greater than 50%. That is, *p* should be more likely true than not, from *S*'s point of view. But this may be too low, or too uniform, a standard for belief, and in any case the proper definition of subjective probability is also controversial. My subjectively rational belief is approximately what William Alston means by deontologically justified belief—"Concepts of Epistemic Justification," in his *Epistemic Justification: Essays in the Theory of Knowledge* (Ithaca: Cornell, 1989), pp. 81-114. Alston's conception is much more elaborately defined, to respond to problems about which my view is indifferent.

I also want to distinguish this private aspect of rationality from public or *objective rationality*, which determines what *anyone* ought to believe, given whatever evidence is *generally* available at the time I realize that this is vague, and I apologize for using the word 'objective' here to mean something like globally intersubjective It might have been better to reserve the term 'objective rationality' for just the formally correct relation of belief to evidence But I am trying to get at what we mean when we say that some belief is rational simpliciter, not expressly relative to one believer or a certain pool of evidence This common usage, in statements like 'it is not rational to believe that the world is flat', seems to require an implicit reference to the pool of evidence to which *we* (that is, some relevant class of normal people) all have some sort of access When we talk about *the* evidence for a belief, as distinct from someone's evidence in particular, this is the pool we mean

So, relative to the premises that Socrates is a man, and that all men are mortal, it is rational (in the pure, formal sense) to conclude that Socrates is mortal If a person *S* at time *t* justifiedly believes the two premises, then it is subjectively rational for *S* at *t* to believe the conclusion Since it is public knowledge that Socrates is in fact a man and that all men are mortal, we can say that it is objectively rational (or, as we often put it, just plain rational, or rational given the evidence) to believe that Socrates is mortal

The Socrates example is one of deductive reasoning, a matter of a logically necessary connection between the evidence and the conclusion drawn But I shall take it for granted that there is such a thing as inductive reasoning as well If we are not to be skeptics about empirical beliefs in general, then we must allow as rational some beliefs that depend on inductive (or perhaps "abductive," or anyway, nondeductive) generalization and prediction Both the nature and the justification of inductive inference are mysterious, but whatever is the right connection between empirical evidence and empirical conclusions, it should apply in both the subjective and the objective case For example, it may be subjectively (inductively) rational for one untutored person in Europe to believe that all swans are white, and for another in Australia to believe that all swans are black, while it is objectively (inductively) rational, that is, rational given the pool of evidence available to people in general, to believe that some are white and some are black

K·P. - 1115

As the products of subjective rationality for two different agents can conflict, so can either one conflict with those of public or objective rationality A certain person at a certain time may be denied important evidence that most other people have, and therefore come to different conclusions from theirs, without committing any errors in reasoning Or an individual may have some extra evidence, beyond what lies before the general public, but which might turn out to be misleading For example, I may have been kidnapped and returned to Earth by aliens, with all of the extraordinary experiences such an event is liable to involve, though nobody else on Earth has any evidence of this fact So only I believe the truth about what happened to me, and everybody else who disagrees is wrong—but that is not their fault Or perhaps it was all an incredibly well-wrought hoax, perpetrated on me by one Steven Spielberg at a cost of $160,000,000, for no reason of which I could be aware In this case, I am the one who is wrong—but it is not *my* fault In the example, we are all reasonable people (except maybe Spielberg), all doing our best with the imperfect evidence at hand This is the model which I wish to apply below to disagreements over religion and many other questions, where it often happens that one side (or each side) considers the other irrational I want to argue that the main source of disagreement in many such questions is that people on the different sides have different total evidence The people on both sides are drawing the conclusions that as individuals they rationally *ought* to draw from *their* evidence Sometimes that evidence is misleading, in which case the conclusions drawn, on one or both sides, turn out to be false And this is not anybody's fault, no one is making a mistake in reasoning Sometimes we all just get stuck

The second main distinction that I want to make is between two levels or orders of evidence, belief, rational belief, and the like, regarding testimony *First-order evidence* is whatever is available to an individual without dependence on the word of others Whatever comes to me directly through my senses, memory, and faculties of inference is my first-order pool of evidence *Second-order evidence* is whatever one can access only indirectly, through reliance on the word of others So if I hear someone say to me that it is raining in Paris, then I have first-order evidence that such-and-such a person has made the sound, 'It is raining in Paris', and I have second-order evidence that it is raining in Paris One's first-order beliefs are then those based entirely on first-order evidence, and one's second-order beliefs are those based to a significant extent on testimony[3] First- and second-order rationality may then be defined as the correct reliance on first- and second-order evidence to form first- and second-order beliefs

---

[3] I am ignoring Wittgensteinian claims about the essentially second-order nature of meaning It does not matter to my arguments here whether or not first-order experiences can be articulated in a purely private language

A common question about second-order, testimonial beliefs is whether they can be reduced to, or rationally derived from, more basic first-order beliefs Hume thought that they can, Thomas Reid and several recent writers[4] have claimed that they cannot This is not a central issue here It is enough for my immediate purposes that testimony be seen as a reasonable form of evidence, derivative or not that the fact that someone says something sometimes counts rationally in favor of someone else believing it Typically, at least, this will happen when the second person has good reason to believe that the first person is a reliable source of information, at least on the topic in question Thus, the efficacy of second-order evidence will seem to depend, in practice if not in principle, on first-order evidence about its sources

I do not claim that this distinction between first and second epistemic orders is at all crisp For example, there are plenty of mixed modes like telescopes, which can rationally be relied upon by individuals only through a justified trust in the craftsmen who made them, their instructions, and so on Even our own sense organs take on this mixed character, once they have been examined and pronounced however good they are by reliable others, which must affect our confidence in using them thereafter But the distinction should be clear enough for present purposes the arguments that I want to present about first- and second-order thinking are not very sensitive to such details

Finally, I need to make it clear that I am talking only about rational or justified belief, as distinct from knowledge [5] We allow ourselves to say that people rationally believe things that they do not know There are three main reasons for this First and most obviously, knowledge implies *truth*, and many justified beliefs are false I may have believed with good reason that my wife was cheating on me, and then discovered (maybe only after I have strangled her) that I had been expertly deceived toward that conclusion by my best friend—so, plainly, I could not have known what was not true to begin with Second, even ignoring the issue of truth, the term 'knowledge' implies greater

---

[4] For two main examples, see C A J Coady, *Testimony A Philosophical Study* (New York Oxford, 1992), and Tyler Burge, "Content Preservation," *Philosophical Review*, CII, 4 (October 1993) 457-88

[5] For convenience, I shall ignore the issue of probabilized beliefs or degrees of belief, and just treat belief as an up-or-down decision

certainty than mere belief I may believe with good reason that Ralph Nader will lose the presidential election in the United States in 2000, and be right, but not know it in the sense of being truly certain Third, knowledge implies *understanding* in a way that simple justified belief does not I may believe with good (second-order) reason that photons exhibit something called wave-particle duality, but I barely understand what this claim means in terms of the theory that explains it—hence, in a way, I do not really know that the claim is true

I am particularly concerned to avoid a side issue over whether there is such a thing as second-order knowledge, in the full sense of the term Plato for one, and John Locke for another, firmly deny that mere testimony can ever provide us with real knowledge [6] But this depends on their strong view of knowledge as something like understanding, not justified true belief Both Plato and Locke do accept that testimony can often be used to transfer mere justified or rational belief (true or otherwise) from one person to another This weaker claim is all that I need for my argument that most people do believe what they rationally ought to believe

II THE SUBJECTIVE RATIONALITY OF FAITH

Sometimes, the total pool of evidence we have will make it rational for us to override our own tentative, first-order conclusion about something in favor of someone else's expressed opinion This will typically occur when we have sufficient reason to believe that the other person or group has a more reliable overall view of the question than we do To the extent that another person is an eyewitness, or an expert, and we are not, it is likely to be rational for us to forego our own first-order beliefs in favor of theirs Even when our own view is privileged in some way, we should sometimes still defer to the opinions of others For example, when I think, based on my own direct experiences, that I am having a heart attack, but then my doctor checks it out externally and tells me no, that I am only suffering from something called "acid reflux," then I should probably just abandon my own prior opinion and adopt his in its place This is because he is much more likely overall to be right, under the circumstances, than I am In general, if we are rational, we should believe whatever is most likely to be true, given all the evidence we have It is, in principle, irrelevant whether the initial source of this belief is our own first-order judgment or someone else's In a conflict of belief, if we are justified in believing that someone else is more likely to be right on the question than we are, then we should defer to their opinion As a strictly subjectively rational matter, we should "stick to our guns" when faced with controversy only if we are justified in believing that we are the ones most likely to be right[7]

---

[6] Here is Plato in the *Theaetetus* 201c "suppose a jury has been justly persuaded of some matter which only an eye-witness could know, and which cannot otherwise be known, suppose they come to their decision upon hearsay, forming a true judgment then they have decided the case without knowledge, but, granted they did their job well, being correctly persuaded "—M J Levett, trans (Indianapolis Hackett, 1992), p 80 See also M F Burnyeat, "Socrates and the Jury Paradoxes in Plato's Distinction between Knowledge and True Belief," *Proceedings of the Aristotelian Society*, Supplementary Volume LIV (1980) 173-91 John Locke agrees, in Book IV, Chapter 14 of his *Essay Concerning Human Understanding*, Peter Nidditch, ed (New York Oxford, 1975), pp 655-56, he uses the term 'probability' to cover all justified belief short of knowledge "Probability then, being to supply the defect of our Knowledge, and to guide us where that fails, is always conversant about Propositions, whereof we have no certainty, but only some inducements to receive them for true The grounds of it are, in short, these two following First, the conformity of any thing with our own Knowledge, Observation, and Experience Secondly, The Testimony of others, vouching their Observation and Experience "

I should make it very clear that I am not suggesting that one can think somebody else's thoughts When I adopt through deference somebody else's opinion, the belief that I end up with is still definitely mine I am not rationally out of the picture just because I have deferred to the beliefs of another I am still drawing the best conclusions that I can from the total evidence available to me, which of course includes my own direct experiences and reasoning based on them, but also includes my experiences of other sources of information besides myself At the end of the day, my coming to the best *overall* conclusion may entail giving up my own *first-order* conclusion But it is still me thinking, from my own traditional, first-person point of view So my claim implies no form of "group-think" social epistemology, that is, one that takes social groups as basic subjects of belief There is no relativism, or social constructivism, or anything of the sort involved in what I am saying I am speaking only of the total evidence that each individual has for his own beliefs, and claiming that this evidence can sometimes force a reasonable person into epistemic conformity with others around him

Still, the principle of rational deference for which I am arguing is very strong It does imply that there are cases in which, if the practical question is only what one should believe at the moment, one might as well simply shut down one's own epistemic faculties, and rely reflexively on someone else instead In general, it says that person S_1

---

[7] John Hardwig argues along similar lines for a somewhat more restricted principle of deference, to the effect that laymen can have good reason to defer to experts, especially in matters of science He declines to count this reason as a kind of evidence See his "Epistemic Dependence," this JOURNAL, LXXXII, 7 (July 1985) 335-49

ought to defer to person S_2 with respect to proposition p just in case S_1 has sufficient reason to believe that S_2 is more likely than S_1 to be right about the truth or falsity of p But it can happen that S_1 finds out, or at least comes rationally to believe, that S_2 is *always* more likely to be right than S_1, in which case it looks like S_1 should just adopt S_2 as an authority, and stop trying to figure things out for himself This may be true for a certain class of mentally retarded people, for example, who come correctly to believe that their own opinions (that is, their own first-order beliefs) about most things are simply not to be trusted, and they should rely instead on some set of trustworthy, nonretarded people around them A larger class is children Almost all small children, at least up to five or six years old, including very intelligent ones, are in this position with respect to their parents The children know enough to know that they are far less reliable overall than are their parents, hence that they are rationally best off (though perhaps not best off overall) accepting whatever their parents say, on almost every subject[8] Of course, ordinary parents are not *really* especially reliable on every subject, including theology, physics, and so on But the young child is in no position to make such distinctions From his point of view (that is, given all of the evidence that he has accumulated so far), whatever his parents tell him about virtually anything is more likely to be true than are the products of his own, separate efforts

Consider an ordinary child's belief in the existence of Santa Claus It may seem that such children must be making some kind of juvenile mistake in reasoning to come to such a poorly evidenced conclusion But from the subjectively rational point of view, there is usually nothing wrong with them at all They are not making any mistakes, given the evidence they actually have They believe in Santa Claus because their parents have told them that he exists, and their parents are the most reliable people that they know Even if they find the first-order case for Santa's existence pretty unconvincing on its face, and even if their older brothers and sisters have explained repeatedly to them what suckers they are being, the fact remains that most kids still have greater total reason to believe what their parents are telling them, however things otherwise seem It is, ordinarily, only once the parents themselves have relented, that a young child gives up this belief, however suspicious or "conflicted" he has become in the meantime.

---

[8] There are other important features of a person's education beyond coming up with the most probable beliefs It is commonly in one's long-term interest to develop one's own critical faculties as well, for example, even at some cost to the correctness of their short-term beliefs Perhaps this helps explain why children go through phases of unreasonable defiance, starting at a very young age

I am not trying to pose as a psychologist. It may be that the best causal explanation of why children tend to accept the beliefs of their parents involves other forces in addition to the kind of reasonable inference that I have described. My point is not that they come to hold such beliefs entirely because they are rational. What I am saying is simply that many such beliefs *are* rational, and surely this is an important aspect of the total explanation for their being held.

This kind of reasonable belief in Santa Claus can be clearly distinguished from a case presented annually in the comic strip *Peanuts*, where the character Linus has his own, idiosyncratic belief in what he calls the Great Pumpkin, on the basis of no evidence whatsoever, either first- or second-order. Every Halloween, he expects this thing to rise out of a local pumpkin patch and bring presents to all of the children in the world. And every year, it does not happen, and he sees that it does not happen, and he keeps believing in the Great Pumpkin anyway, all by himself. I say that this extreme "fideist" Linus is subjectively irrational, where the ordinary Santa Claus believer is not.

My point applies to true as well as false beliefs, of course. For example, I believe that the last Tsar of Russia was a man named Nicholas, not because I have any first-hand evidence for this fact, but just on other people's say-so. If I refused to believe in this Nicholas, just because I have no direct evidence of his existence, then, given what I know about the reliability of many of my sources, I would be making a serious error in reasoning. Or, if I believed on the same evidence that there were two such people, Nicholas and another one named Freddie, I would also be committing an error. In such matters—and there are many such matters—we should believe no more or less than what we have been told.

Now I want to apply the same basic analysis to religion (more specifically, religious *faith*) in general.[9] The prevailing view is that religious faith is an irrational (or just nonrational) alternative to belief based on evidence. Explanations of other people's faith vary from hope, fear of death, and the like to political "brainwashing," but the common thread is that believers are persuaded by essentially psychological or social, not epistemic, forces.[10] But I say that, if someone grows up in a society in which his parents and everybody else believe in reincarnation, for example, then, if he is rational, he will defer to all of his elders and believe in reincarnation, too. He may have no direct, first-order evidence for this view (though he may or may not have been taught that some things *count* as first-order evidence for it), but this makes no practical difference. Given the evidence that he actually has, he must conclude that those around him are much more reliable, at least on big public issues, than he is himself. Therefore, if he wants to believe what is most likely to be true, he ought to accept whatever these trustworthy people, his elders, tell him.[11] If what they say makes little sense to him, then that is too bad—just as it is too bad for me that I cannot make heads or tails of quantum mechanics.

---

[9] Religion and faith are not coextensive. There are plenty of traditional and other second-hand beliefs that have nothing to do with religion, and there are sources of religious belief that are not second-order. For example, many people claim to have their own, autoempirical religious experiences, anywhere from hearing vividly the voice of God, to the vague "oceanic feeling" that Sigmund Freud talks about. There are also some science-like ingredients of religion, including philosophical discussions among theologians. But the thing that really makes religion or religious faith what it is, an institution as opposed to a mere set of doctrines, is its transmission through testimony from one generation of believers to another.

We say that people ought to "think for themselves." But what would that be like in these cases? Should I do my own particle physics? If not, then why should every member of a traditional society do his own metaphysics? How could it be rational for him to evaluate the cosmic situation on a purely first-order basis, and then accept his own restricted conclusions in preference to the unanimous testimony of the most trustworthy people he knows? To think that one can figure this sort of thing out for himself would imply that he believes that his own first-order epistemic machinery is somehow more reliable than everybody else's Unless he is an extremely unusual person, he has no evidence for this at all Even if it strongly *seems* to him that he has good direct evidence for his own, different first-order view, and that everybody else is screwing up somehow, his knowledge of their overall epistemic record and abilities, compared to his own, should surely force him to defer Those few people who refuse to defer under such circumstances can be said to suffer from the second-order form of irrationality which is usually called arrogance The more common attitude toward doctrines that may not seem directly to be right, but are adopted on good grounds of testimony, is what Christians since Augustine often really mean by *faith*, and what I am calling second-order rationality[12]

---

[10] The biologist Richard Dawkins gives a particularly blunt and hostile statement of this view: "I think a case can be made that *faith* is one of the world's great evils, comparable to the smallpox virus but harder to eradicate. Faith, being belief that isn't based on evidence, is the principle vice of any religion. And who, looking at Northern Ireland or the Middle East, can be confident that the brain virus of faith is not exceedingly dangerous?"—"Is Science a Religion?" *The Humanist*, LVII (January-February 1997): 26-29. He goes on to describe ordinary religious education as a form of "mental child abuse." Most opponents of religion are at least more diplomatic.

[11] Alston discusses a similar case (*op. cit.*, pp. 95, 145-46) and argues that a person who grows up inside a culturally "isolated community" cannot be *blamed* for adopting the peculiar traditions of his elders, in part because he will never have considered questioning them. In this weak sense, Alston allows that such people can be "deontologically justified" in their beliefs (his overall point is that such a conception of justification is faulty, precisely because it leads to such cases). The position that I am suggesting here is much stronger: I say that ordinarily, a culturally isolated individual is (from a subjectively rational point of view) positively *obligated* to believe in the local traditions, even if he has considered them quite thoroughly, and even if he directly perceives what is wrong with them. The only possible exception is if the individual has sufficient total reason to believe that he is more reliable on the subject than the sum of his elders.

This is what explains the great stability of religion and other traditional beliefs It is because human beings are usually quite rational in the subjective sense—that is, we make about the best sense that can be made of the total evidence we have, including second-order evidence—that normal people growing up in traditional societies so "uncritically" absorb the views of those who went before them The common prejudice among scientific Westerners, that generation after generation of highly civilized Egyptians, Chinese, Hindus, Incas, Moslems, and the like are all somehow deficient in their reasoning, shows its absurdity, I think, when placed against this reasonable alternative

Here is a major objection, first posed to Augustine's view by Hume, in his famous essay on miracles[13] What if a religious claim is so bizarre that the probability of even unanimous testimony in its favor being true is lower than the probability that everybody else is simply wrong? Hume claims that miracles are, by definition, events of very low probability, since they violate what seem to be the laws of nature To be persuaded by second-order evidence of miracles, then, requires that one assign an even lower probability to the proposition that the testimony in question has somehow misfired, through past deceit, ill-will, undue credulity, mistaken perception, or any other fallibility of our reliance on others Hume asserts, quite reasonably, that in the case of a conflict, the real probability of any human authority's being correct will never be as high as the probability that the usual apparent laws of nature are in force

---

[12] Augustine makes this testimonial notion of faith a central premise of his conversion to Christianity In his *Confessions*, Book VI, Chapter 5, he says "I began to realize that I believed countless things which I had never seen or which had taken place when I was not there to see—so many events in the history of the world, so many facts about places and towns which I had never seen, and so much of what I believed on the word of friends or doctors or various other people Unless we took these things on trust, we should accomplish absolutely nothing in this life Most of all it came home to me how firm and unshakable was the faith which told me who my parents were, because I could never have known this unless I believed what I was told"—R S Pine-Coffin, trans (New York Penguin, 1961), p 117

[13] *An Enquiry concerning Human Understanding* (New York Oxford, 1957), chapter X

From an external point of view, Hume is correct about these probabilities So his argument is valid, objectively speaking But he is not considering every ingredient of the *subjective* probabilities involved, when an individual is actually confronted with a question like this In particular, he leaves out what the individual may justifiedly believe about his own ability to figure such things out Even if the individual has read Hume's argument himself, and finds it very persuasive on its face, he does not know for sure that he has understood it fully If he then shows Hume's essay to his parents, teachers, and the like, and they all say that it does not work, this counts as a powerful, subjective argument against it *for him,* even if he cannot grasp their actual objections

But what if the religious (or other second-hand) belief in question contains, not an improbability, but an outright contradiction? If rationality is supposed to be a useful thing, must not the rational person be allowed to apply autonomously at least some basic tests for coherence? The answer is "yes" and "no " Rationality does plainly require that one's overall view of things be basically coherent But this does not entail that each belief that a person has must make sense *all by itself* to that person, when this conflicts with rationality's demand that one make the best available sense of one's *total* epistemic situation If everything that your elders ever told you had struck you as highly improbable in the first place, then you would indeed have good reason to doubt the next implausible suggestion that you hear from them, based on their previous performance But, to whatever extent they have established their reliability (in the relevant subject matter) to you, you have that much inductive reason to believe whatever next thing they assert, however oddly it strikes you If they have established a degree of relevant reliability that is superior to your own, as is usually the case for parents and their children, then you can easily end up with better overall reason to defer to what they tell you than to believe something else that appears to make far better sense If such a situation causes some local inconsistencies to crop up within one's set of beliefs, this is unfortunate, but often unavoidable

Think about the problem of evil You grow up believing, on the basis of the testimony of your parents and other adults, in an all-powerful, all-knowing, and completely benevolent God One day it occurs to you that the existence of such a God is frankly contradicted by the pointless suffering experienced by millions of innocent people (and other creatures) every day Now, what should you think? I say that you are faced with a choice that is, in principle, still pretty simple Should you continue to trust your elders in this matter, or should you trust yourself instead? Admittedly, you have found yourself in a certain psychological state—the state of being persuaded by an argument that a certain complex proposition is incoherent But is this a reliable indication of the truth? Has something suddenly made you an expert on what is and is not actually incoherent, as distinct from what merely seems to be incoherent? Suppose that your parents, priests, and others now tell you that they are aware of your argument, and assure you that it can effectively be countered (through a very subtle rejoinder, or perhaps through some other process that goes "beyond" argumentation) As long as you understand that your own faculty of reason is not a superior indicator of necessities and possibilities, then you will have little more total reason not to defer to your parents now than you did before the problem occurred to you This is so, moreover, even if you turn out to have been right, and the problem of evil is in fact objectively devastating to traditional conceptions of God—indeed, even if you "know" that it is so, in the sense that you are aware of the correct first-order path to that conclusion

The same point holds outside of religious or traditional beliefs I have a friend—a philosopher, not a physicist—who claims to have proven that relativity theory is incoherent He may, for all I know, be right But I am sure that this person is not justified (in the subjective sense) in holding this as a belief, if he really does It almost does not matter whether relativity actually is incoherent, or how good this fellow's argument is, objectively speaking Once he has shown it to the experts, and they have turned it down, he simply must defer to them, if he is to be rational about it I am in roughly the same situation myself, with respect to the coherence of quantum mechanics—except I realize that I am in no position to judge the matter, so I just accept the experts' view that this theory is correct It seems self-contradictory to me, but that is not sufficient reason for me to disbelieve a theory, unless it is so frankly incoherent that I cannot even frame it in my mind And even then, it may be rational for me to accept that such-and-such a theory, which I cannot even mentally express, is nevertheless probably true

There is indeed a limit to how crazy one's beliefs can be, but it is a pragmatic, not an a priori one If I believed that all amounts of money are the same, for example, then I would likely fail in life from that belief If I believed that hydrochloric acid is health giving and delicious, then I might well die Beliefs that seriously interfere with the conduct of ordinary life are naturally hard to establish in the first place, and will be equally hard to maintain in the face of a recalcitrant world. This is why many of the most bizarre-seeming tenets of other people's religions have to do with primal history, the afterlife, and other matters outside the scope of daily existence. Still, it is instructive to remember what some people actually put themselves through in terms of daily inconvenience, not to mention some extreme forms of self-denial and even self-mutilation, in order to live consistently with their traditional beliefs.

As John Rawls has said, the proximate goal of the exercise of reason is a kind of equilibrium. Each thing that a rational person experiences, first-hand or second-hand, must be processed as a piece of evidence about the world, factoring in whatever one knows about the reliability of its source. In order to make the best sense of the whole collection, one may well have to sacrifice something of the best sense to be made of some parts. What we seem to perceive with our senses usually counts for a lot, because we have good reason (much of it testimonial) to believe that our senses, especially our eyes, are very reliable. What we can figure out by pure, first-order reason also counts a great deal, for similar reasons, but not necessarily decisively in any case where there is contrary evidence as well. And testimony from any source not known to be deceptive also counts—sometimes less than other kinds of evidence, depending on one's total situation and experience with other people, but sometimes more.

Hence my point about traditional societies: when available testimony on some point is truly unanimous, its epistemic weight can simply overwhelm the other resources, a priori and otherwise, of the individual, rational mind. The most intelligent people in the world may end up believing almost anything, if it is part of an overall theory of the world that makes the best sense out of each individual's total experience, including everything that other people say to them. Religious faith is not, then, just an irrational "opiate of the masses." It is, essentially, the social realization of second-order rationality. These beliefs may have arisen *initially* through people's desire to have a comforting story about death, to control others, and so on—or perhaps some people really witnessed miracles. But once these beliefs have been established, no further "backwardness" of any kind is required for their maintenance, and no political sort of repression, either. Once the process takes sufficient hold in any society, it will tend to persist indefinitely, from each generation to the next, as new members are raised into it. Of course, it is bound to happen that some people grow up deferring to beliefs that they do not find very plausible, or even fail entirely to understand, like many Catholics do not understand the doctrines of the trinity and transubstantiation.

Such obscure beliefs might possibly be viewed as errors that have gotten entrenched, and that the epistemic community simply has no way of correcting from inside[14] I am not denying that it is a bad thing when mistakes or inconsistencies get passed on like this My point is that when this happens, it is not because the people involved are irrational, in the subjective sense It is because they are rational

Unfortunately, too, a certain amount of flat-out intellectual repression does seem to make sense from within a traditional society If all the members of a community are forced by rationality to adopt some belief, due to the great epistemic weight of testimony in its favor, then anyone who attacks that belief is ipso facto behaving unreasonably It is not just raw power, then, which argues for the suppression of opposing views within traditional societies Since the dissenting views themselves are rationally wholly unbelievable (though not for intrinsic reasons), it is hard to see from inside such a community why reasonable people should have to put up with them—especially in legislatures, schools, and other public theaters, where everyone agrees that reason ought to rule Even in a nontraditional, modern society like ours, such repression may seem appropriate where unanimity is nearly total For example, those who now allege that the Holocaust never took place are (and arguably should be) denied nearly all forms of public support for the dissemination of their view, not simply because the view is false, or because it is nasty, but because such a view can only be held *irrationally*, in light of the great mass of physical and especially testimonial evidence to the contrary

Here is a second objection, or a possible restriction My account of rational deference so far assumes something close to unanimous testimony on the part of one's elders or one's whole community But many religious and other traditional believers maintain their views despite conditions of controversy, even as members of tiny minority groups Can at least this kind of deference—to local authorities, in the face of known disagreement from others—be counted as irrational? I say perhaps, but not necessarily It will depend on the immediate epistemic conditions for the individual believer For example, if a child is raised by evidently highly trustworthy parents and others whom he knows, both to accept some minority belief and strongly to distrust the majority that disagrees, then that person's belief may be rationally stable for a long time—perhaps his whole life if he is sufficiently "protected" from the outside world This is certainly a common pattern with minority religious groups like the Amish or Hasidic Jews, or other fairly self-contained epistemic communities. But there is also a tendency, perhaps a stronger one in very mixed societies like ours, for such minority beliefs to break down (that is, to stop being rationally justified) under the pressure of conflicting testimony from "outsiders" who have nevertheless shown themselves to be reliable sources—at least, not too much less reliable than the believer's original community. Travel, military service, and college education have long served this dissolusive role, by confronting previously sheltered individuals with evidence that many nonbelievers can in fact be trusted.

---

[14] The process can be interrupted by an external force, of course, such as a conquering army But even then, traditional beliefs are terribly hard to extinguish, and frequently outlast, or even absorb, such conquerors

Descartes, experienced himself in all three modes of outside contact, raises this issue of conflicting testimony in the *Discourse.* He claims that such conflicts of authority tend to cancel out the total weight of testimony in those situations, throwing the individual back on his own resources.[15] For example, if a third of my elders say that Odin is the king of the gods, and another third say that it is Zeus, and the last third say that there are no gods at all, then obviously I cannot take all of these propositions equally on faith. Let me then just ignore the whole lot of them, and do what I can to figure the situation out for myself. Perhaps I shall end up agreeing with one side in the ongoing dispute, or perhaps I shall come up with my own new theory, for example, that there is only one god, or that the gods are democratic. In any case, according to Descartes, I should rationally come to believe whatever theory makes the most sense to me on first-order grounds, since the pool of second-order information I received was effectively nullified by the contradictions it contains.

I disagree with Descartes about this.[16] I think that the rationally correct procedure is to assemble all the candidate theories one can find, and then place the greatest faith in the one most likely to be true, all things considered. One's own first-order view of things is never more than one of these choices. And when it is opposed to the unanimous testimony of his elders, one's own view is almost always, as

---

[15] See *Discourse on the Method* in *The Philosophical Writings of Descartes,* Volume I, John Cottingham, Robert Stoothoff, and Dugald Murdoch, eds. and trans. (New York: Cambridge, 1988), part one. Also at the beginning of the *Meditations,* he says "considering how many diverse opinions learned men may maintain on a single question—even though it is impossible for more than one to be true—I held as well-nigh false everything that was merely probable"—*The Philosophical Writings of Descartes,* Volume I, p. 115. Descartes sees the need for a first-order criterion of rational belief (ultimately clear and distinct perception) as the only alternative to skepticism.

[16] But I do agree with Hardwig (*op. cit.,* p. 343).

I have said, less likely to be true But this is equally the case if one's own view is opposed by the *divided* testimony of his elders, because he is still less likely to be right than any of them Here is a simple example Plato says *X* about some philosophical issue, and Aristotle says *Y* The best thing I can figure out on my own is *Z* Now, what should I believe? Certainly not *Z*, unless I have good reason to think that I am as reliable a philosopher as either of the other two From a strictly rational point of view, I should conclude that the correct answer is likely to be *X* or *Y* (or something else that I am not aware of), and acknowledge that I am not in a position to decide between them My own view may be interesting to me, and I may gain in various ways from the exercise required to produce it, but I am not in a position to *judge* that my own, personal opinion has any greater virtue than its plausibility to me

The inductive conformism that I am recommending does have some limits I do not wish to suggest that a person should automatically believe whatever the mere majority around him believe If more philosophers believe in theory *A* than theory *B*, this may make theory *A* somewhat more likely to be true, ceteris paribus But this relative likelihood does not itself suffice to meet any reasonable standard of belief Only when there is an established consensus on an issue is it normally reasonable for someone to accept the judgment of others without further question But at the same time, it is only in the unlikely case of a total breakdown of testimonial authority that one is ever truly thrown back on his own epistemic resources If the disagreement among your elders reached the point where you had reason to judge that none of them was more reliable than yourself, then your own opinions could properly be seen as at least equally worthy of belief, compared to the theories of others But again, it is only ultimately rational for you to prefer your own opinions if you have reason to prefer yourself as a source of opinions If no one in the world, including you, stands out as a reliable authority on the relevant questions, then the rationally best thing to believe is nothing at all

At the same time, there is undeniably a certain intellectual freedom that results from such an absence of coherent authority If, as suggested below, it is natural for people to want a picture of the world that is reasonably complete, at least where action is likely to be required, then even in cases where a suspension of belief is rationally favored, the desire to believe may take control psychologically This may explain why first-order reasoning seems naturally to thrive in those complex epistemic situations where traditional authority has broken down, and new beliefs, including quite bizarre ones, seem to arise and flourish with only the flimsiest objective evidence in their favor People anxious for a theory, but deprived of all authoritative guidance, might simply grab whatever is most psychologically attractive, on the rationally nondecisive grounds of first-order plausibility to themselves

In ordinary intellectual life, where authority is less than absolute but holds some sway, first- and second-order justifications for belief are all mixed up Our belief in our own intellectual powers is largely conditioned on our accepting the authority of various other people who tell us how reliable they think we are, and how reliable the various other authorities are This judgment, in turn, affects our estimation of the reliability of each of them, and everybody else For us to develop the best possible picture of the world, our conceptions of our own intellectual strengths and weaknesses must come into equilibrium with our beliefs about everybody else's, and in turn with our beliefs about the actual objects of everyone's opinions

This is why it is almost never sufficient to justify upholding a dissenting opinion, merely to come up with a coherent first-order explanation as to why the orthodox opinion is wrong Someone who believes that everyone else is a devil, for example, has a consistent view, which accounts coherently for everyone else's testimony to the contrary Rationality requires, not just that one construct *some* coherent account of one's entire evidential situation, but also that one make the *best* total sense of it And it is highly unusual for the best overall theory to be one in which first-order evidence outweighs anything like unanimous testimony to the contrary

III THE OBJECTIVE RATIONALITY OF SCIENCE

We tend these days, as I have noted, to think of empirical science and associated attitudes toward discovery and truth as normal, and of religious belief and other kinds of "folk theory" as representing a failure of fundamentally scientific rationality But the truth is the other way around Science is rare in history, for the reason that it requires, in effect, the *suspension* of ordinary individual rationality, in the ultimate pursuit of knowledge that is largely unavailable to its immediate discoverers In general, for each basic scientist or other intellectual explorer, the probability of his own new theories turning out to be true will be far lower than the probability that the received view of the time, or some other person's innovation, will be proven right The young child who announces, on the basis of his own observations and deductions, that Santa Claus does not exist, despite his parents having insisted that he does, may be said to be to be right in the sense that what he says is true, and is backed up with what turn out to be good first-order reasons But he is also wrong, in the sense that he has no justification for preferring his own overall judgment to that of his parents[17] The same goes for scientists and other intellectuals who wish to challenge the beliefs of their own authorities For science to work requires that scientists *ignore* the epistemic weight of much of what they hear from those around them, since a rational respect for such testimony can place the subjective probability of any very new hypothesis at close to zero

Here is an example, simplified a bit to illustrate my point Galileo is a hero of science, but may not have been an altogether reasonable man He certainly defied his elders (if the church authorities of his day can be counted as such), not just by disagreeing with received opinions, but also by denying that religious or social authority should carry any weight in matters of science Galileo had the telescope, of course, and access to other recent developments, as well as the fruits of his own reasoning But so did his opponents, who never seriously denied that astronomical *appearances* were as he claimed The main issue between them was one of how to decide, and who gets to decide, what these appearances amount to in the greater scheme of things And Galileo was, for much of his working life, openly contemptuous of the "ignorant and superstitious" beliefs of the authorities around him He was ultimately brought to trial by the church, gently imprisoned, and required to renounce his unorthodox views Now, the common perception is undoubtedly correct, that Galileo was the premier founder of the modern scientific method, and for this he certainly deserves our thanks He also authored many particular ideas which turned out to be right, or at least important improvements on the going theories of his day So we can say that his opinions were rational ones, in the objective sense But was he justified subjectively, in believing these new theories himself? There are grounds to think that he was not, because he never had sufficient overall reason to prefer his own approach to the received view of the Church If he thought that he did, this would have required him also to believe that he, Galileo, was a better judge of the entire situation, encompassing both science and religion, than were the church authorities who argued that reliable scriptures ruled out certain of his physical hypotheses And it is not clear how he could have had sufficient reason to believe this claim about himself (even if it was, objectively speaking, largely true) He must have known that he was very bright But unless he believed justifiedly that he was brighter than everybody else, or privileged in some other general way beyond the resources of his elders, there was no adequate internal reason for him to prefer his own authority, as it were, to theirs We are glad that he did, of course, because he turned out to be right, at least approximately, both on substance and on scientific method But he himself could not have known that Far from being a martyr to reason, as he is usually portrayed, he was a man whose total epistemic situation made it rational for him to recant when he was finally ordered to do so [18]

---

[17] Again, I am only speaking of purely rational justification There may well be other good reasons, perhaps developmental ones, for children to be inclined to reject their parents' testimony

In the practice of science, subjective and objective rationality part company The fact that radical inquiry of the sort that Galileo practiced is ordinarily irrational for individual thinkers makes it unlikely that science should ever emerge inside of a traditional society But the fact that empirical science *works* in the long run—that is, leads to more true beliefs, providing for better technology, weapons, and so on—makes it hard to kill once it gets started It spreads, not because it makes the most sense for the individuals who practice it, but because the societies in which it has taken hold tend to be more successful than those in which it has not A scientist devotes his career to the development of some new theory Is it likely to be true? No Each individual scientist is probably just wasting his time, if the goal is merely to increase his own stock of true beliefs [19] But there are many scientists working at once, on many hypotheses If any of them prove to be correct, then all of the consumers of the scientific product benefit, albeit at the expense of unsuccessful scientists, to the extent that the latter have convinced themselves of theories that will turn out to be false [20] Thus, scientific reasoning appears to involve a kind of *epistemic altruism* on the part of individual scientists, intentional or not [21] By refusing to defer to others when it is subjectively rational for them to do so, they sacrifice the greatest likelihood of having true personal beliefs, for the sake of producing better beliefs for everybody in the long run [22]

---

[18] This is only the first layer of a full analysis of Galileo's situation In real life, Galileo was not truly alone, and his opponents were hardly unanimous

[19] Though one's *knowledge*, in the sense of understanding, is again a different matter The practice of philosophy or science may well enhance one's understanding of the world, even while diminishing one's chance of being merely right A "defiant child" like Karl Marx or Freud, for example, might be seen as having gained a deeper overall appreciation of his subject, while being less correct than prior common sense about most points of fact Such comparisons are, of course, impossible to quantify

[20] What I am saying obviously applies more directly to what Thomas Kuhn calls revolutionary science than to his "normal" science It is not, in general, the incremental advances, in accordance with accepted paradigms, that require suspended rationality on the part of their authors, it is big leaps of genius, in the face of direct opposition from one's elders In ordinary life, professional scientists and other academics are in fact encouraged to function semi-autonomously, on the basis of their mastery of authorized procedures And this is not so different from the authority structure of the Church in Galileo's time, which also allowed for widespread scholarship and speculation, in that final authority still resides very firmly at the top of each profession In the recent controversy over "cold fusion," for

This is true not just for empirical science Intellectual progress in general results from the same kind of altruism New ideas in ethics, cultural life, and even art must be launched initially in the face of such great second-order evidence to the contrary, that it makes little sense for their proponents to have any confidence at all in such hypotheses The autonomous thinkers themselves are likely to lose, from a strictly rational point of view, whenever they actually believe in their own new ideas But their successors, and their societies as a whole, gain much more when one of their hypotheses (of whatever sort) proves to be true or useful [23]

---

example, a small group of highly credentialed university scientists was decisively rebuked by professional organizations (including funding authorities) when it was decided that they had not correctly followed authorized procedures For more on this case, and a far more detailed discussion of how authority actually works in science, see Philip Kitcher, "Authority, Deference, and the Role of Individual Reason," in Ernan McMullin, ed *The Social Dimensions of Science* (South Bend, IN Notre Dame, 1992), pp 244-71

[21] It is controversial among both psychologists and biologists whether altruism is truly possible Some (for example, Elliott Sober and David Sloan Wilson) believe (following Charles Darwin) that altruism can evolve through the mechanism of group selection, where groups of cooperatively self-sacrificing organisms are able to outperform other groups comprising individually fitter, but less mutually cooperative ones—*Unto Others The Evolution and Psychology of Unselfish Behavior* (Cambridge Harvard, 1998) This is an inviting model for the notion of objective rationality that I am suggesting is the basis of science, but my point does not depend on such a theory being right I am only claiming that science tends to progress against the purely *epistemic* self-interest of individual participants, in favor of the epistemic interests of others As with other species, it is not required that this type of altruistic behavior in humans be recognized as such by its performers In any case, scientists are usually compensated in a variety of nonepistemic ways for their efforts, so it is arguably still in their *overall* self-interest to do what they do

[22] I reiterate that I do not mean to imply that it is only whole societies (as opposed to individuals) that can know (or justifiedly believe) things, or any such radical thesis When I say that a society gains in knowledge from the efforts of researchers who cannot know whether their hypotheses will be proven correct, I simply mean that future individuals in that society can know what their forebears could not, once an appropriate consensus is reached

[23] Not that science and similar intellectual activities are unstoppable once they get established Many times, societies have reached philosophical or technological plateaus, as in ancient Rome and China or medieval Japan, where a period of exploration gives way to some new set of orthodoxies Even in the modern West, various forces periodically press for slower change, consolidation, or even a return to "basics" of various sorts

It does appear that people naturally prefer believing something to believing nothing As Hume says "Nature, by an absolute and uncontrollable necessity has determin'd us to judge as well as to breathe and feel "[24] It is easy to see how the need for action in a largely hostile world entails a need for some kind of theory of that world, even if the only available theories are unjustified Hence, rationality itself may be of limited survival value A creature who found it very easy to suspend judgment whenever evidence was incomplete or contradictory would probably be less fit for survival than another who was more ready to jump to conclusions—for example, in the matter of whether one was about to be eaten by a tiger

But it is still not absolutely necessary for philosophers and scientists to believe their own theories, or believe anything at all on a given issue [25] Intellectuals of all sorts can distinguish between the role of a creator or advocate of ideas, and the role of a disinterested judge of the truth What is a good means of intellectual production, and what are good grounds for belief, are really very different questions Philosophers and scientists are trained to produce new intellectual material that can be ground up and examined by others, and to engage in controversy in a spirit of healthy competition But a fully rational person ought to withhold belief from hypotheses that have not yet been established, or at least made probable, regardless of his own personal role in generating or defending those propositions To think that one's own products are especially deserving of assent—this is where a potentially useful, speculative posture turns into hubris It is always possible to doubt, at least "metaphysically," even those thoughts which one finds most psychologically compelling For a normal person can resist *assenting* to anything he knows is improbable, as we do when we find out that we are hallucinating, or that the proverbial stick in the water really is straight Proper internal use of the phrase 'it seems' should be adequate, though it must be sincere

---

[24] *A Treatise of Human Nature* (New York Oxford, 1967), I IV I, p 183 Here he means to judge in general, not to make any particular judgments But he also clearly states that beliefs about particular objects can sometimes be "extorted by the present view of the object," though he says further that this implies no "dogmatical spirit" or conceit (*ibid*, I IV VII, p 274)

[25] Most likely, we have evolved a differential mechanism, leaving us best able rationally to suspend judgment in those matters where action is least likely to be required This may explain why most of us seem to need a full set of political opinions, but few of us feel compelled to have any belief as to whether there is life on other planets It may also explain why philosophers and other sages are expected to withdraw, to some extent, from the concerns of ordinary life In order for our rationality to reign in all matters, this natural need to believe, based on expected needs for action, must be to some extent suppressed

For example, a raven may look totally, convincingly black to me, but nothing forces me to conclude from this that the raven really is black, since I always have the option of concluding, in the face of any sort of evidence to the contrary, that it merely seems to be black

Present-day physicists have the best possible attitude, I think, in not expecting themselves, individually, to be the ones who will turn out to be right They have seen so many large and small "paradigm shifts" occur, even within the span of a single career, that they have learned to be quite skeptical of current theories Still, they engage in their work with the zeal of people very close to achieving all that they desire Their suspension of personal judgment does not dampen scientific enthusiasm for them, much as this would have surprised the ancient skeptics[26] Perhaps the difference is that individual physicists do expect to see new pieces of the truth keep emerging through the institutional inquiry in which they play a part, so they consider their suspense to be essentially temporary These scientists can see themselves both as producers of relatively raw ideas, and as potential consumers of a largely finished product

As for the ordinary, modern consumers of ideas, they must respect the great success of science overall, but also pay attention to the frequent overthrow of current theory Their most reasonable attitude bets heavily on science in the long run, but against the view that this or that particular theory is the whole truth of the matter Many highly educated people come from a partially traditional background as well, and this often creates special epistemic problems for them—for example, in the apparent conflicts between biblical creationism and modern geological and evolutionary theory It may be that, as between scientific and religious authority, some such individuals will have enough total evidence to "fall" one way or the other, while others are left with no clear, reasonable choice at all, and may suffer greatly from the cognitive (as well as social) strain that this creates Ironically, widespread attacks on the rationality of faith will tend to push some people out of this uncertain state by second-order means alone, while pushing others into it

IV CONCLUSION

Here is a final objection to my view It can be claimed that I have no business asserting my thesis about controversial beliefs, given the fact that the thesis itself is novel, and bound to be controversial if discussed at all. Thus I am caught in an apparent paradox: if what I say is true, then by my own arguments I ought not to believe it. It does not make sense to be the only conformist, as it were, in town. But there is really no paradox here, since I desire to assert my thesis only hypothetically—not as a report of my own firm beliefs, but as a mere, suggested avenue of inquiry for others to pursue and evaluate. I have first-order reasons to believe in my conclusions, which I have given above. If I did not find the view pretty convincing myself, I would not have bothered writing it all down. But I really do not know what other philosophers are going to think about these arguments in the long run. If there are glaring errors, I may find that out pretty soon. But if my thesis is true, I doubt that I shall ever know this, because most other philosophers will not agree very quickly, and I already know that my own opinion in philosophy, while modestly reliable as such things go, is not *especially* to be trusted.

---

[26] For example, Kip S Thorne talks about the bets he makes with Stephen Hawking over controversial physical hypotheses—*Black Holes and Time Warps Einstein's Outrageous Legacy* (New York Norton, 1994), p 521 He does not automatically bet on the truth of his own work

Perhaps the only point on which there is a true consensus among philosophers is the principle of epistemic self-reliance itself—which is what I have been questioning. But even if I were attacking this principle head-on (for example, by arguing for a straight, majoritarian conformism), I would still be obeying it, since this attack would be my own idea.[27] In any case, it has not been my purpose to attack the idea of intellectual autonomy itself. Instead, I have been trying to say what it really amounts to, and how it is best to be employed—that is, in the experimental, "brainstorming" spirit of contemporary physics. So my position is a little awkward, perhaps, but no more paradoxical than that of any other intellectual who is working with the proper attitude toward his own, tentative results.

THEODORE J. EVERETT

State University of New York/Geneseo

---

[27] It is more common to encounter the implicit conformist, who unquestioningly asserts the principle of "thinking for yourself," just because everybody else does. The *locus classicus* for this point is the famous balcony scene in Monty Python's film, *The Life of Brian*, where the reluctant savior Brian commands his followers to think for themselves, and they shout their agreement in unison.

BOOK REVIEWS

Unto Others The Evolution and Psychology of Unselfish Behavior ELLIOTT SOBER and DAVID SLOAN WILSON Cambridge Harvard University Press, 1998 394 p Cloth $33 50, paper $17 95

Unto Others is an intellectual joyride, by turns speculative and pioneering, cautious and carefully argued, scholarly and judicious, self-righteous and strident, generous and evenhanded, and defensive and partisan in advancing its central themes the undeserved neglect of group selection as an important mechanism of evolutionary change making possible the evolution of altruistic traits (chapters 1-3), the special significance of this mechanism for the evolutionary history of human beings (chapters 4-5), and the evolutionary considerations favoring the inclusion of altruistic impulses among our ultimate psychological motivations (chapters 6-10)

Elliott Sober and David Sloan Wilson have been putting group selection's best foot forward for quite some time now, and they effectively confront a number of confusions about the issue, most importantly, what they call the *averaging fallacy* defining altruism and group selection out of existence by averaging the fitnesses of genes, traits, organisms, and the like across the various environmental circumstances they encounter (including membership in groups with varying proportions of altruistic members), noting that it is the entities with the highest *average* fitness which evolve, and ignoring the level of biological organization at which the relevant selective processes actually take place Using now-familiar examples (disease virulence, female-biased sex ratios, meiotic drive, and so on), they show how selective pressures can coexist and even conflict at different levels of biological organization most important for them, even an altruistic trait conferring a selective disadvantage upon its bearers within each isolated subgroup of a population will increase in frequency in the global population through a competing process of selection at the level of groups themselves by the operation of Simpson's paradox—here that groups containing higher proportions of altruists contribute more offspring to the next generation of the total population even as the frequency of altruists declines within each group—*if* the group benefit is sufficiently large relative to the individual cost, *if* there is sufficient variation in the proportions of altruists present in the various groups, and (most awkwardly) *if* the population is structured into isolated groups that mix and reassort

0022-362X/01/9801/43-47

(again with sufficient variation in the proportion of altruists within each group) or otherwise compete in the formation of new groups It is disappointing that Sober and Wilson do not provide a wealth of new empirical examples in which these conditions are clearly realized, however (especially since they suggest that such examples abound once our conceptual blinders are removed), because even their detailed and elegant discussions of these familiar examples (and convincing criticisms of earlier analyses of the conditions required for group selection by Sewall Wright and John Maynard Smith) do little to rebut the standard assessment of their position among evolutionary biologists that defenders have shown selection at the level of groups to be a theoretical possibility, but one requiring such unusual conditions as to be rare and comparatively unimportant in nature Of course, part of their agenda is to advertise even this much progress for group selection to scholars outside evolutionary biology

Furthermore, Sober and Wilson do offer an implicit reply to this appraisal, arguing that accounts of the evolution of altruistic traits traditionally considered widely applicable alternatives to group selection—inclusive fitness, evolutionary game theory, and selfish gene theory—simply offer different *perspectives* on the *very same process* of group selection described by their own multilevel selection account This conclusion depends, however, on their controversial conception of a "group" as absolutely any set of individuals whose interactions affect each other's fitnesses, a proposal which has led many reviewers to suggest (mistakenly, I think) that the only issues here are semantic and that the authors try to win an argument over group selection simply by redefining it While Sober and Wilson are right to insist that evolutionarily significant groups are defined by the pattern of interactions among individuals they make possible (rather than spatial or temporal boundaries), their bid for generality nonetheless rests upon supposing that this liberal conception of group selection will be a helpful or perspicuous way to think about the causes of evolutionary change in an implausibly broad range of cases They invite us, for example, to conceive of a population randomly pairing up for rounds of prisoner's dilemma as a single global population that is repeatedly divided into many ephemeral groups of size two and in which group selection runs rampant while cooperative behavior is exploited and therefore selectively disadvantageous within mixed groups (pairs) of cooperators and defectors, groups with higher proportions of cooperators (100% versus 50% versus 0%) ultimately contribute more offspring to the global population (note that this is true of groups containing 50% altruists (that is, one altruist) only because of the greater productivity of the *selfish* member) Although the numbers can be correctly crunched under this description, our understanding of why some strategies evolve and others do not is surely better advanced by the game theorist's usual account, that is, by noting the payoff to each strategy against each potential interactor and the rate at which each is encountered—in a word, by calculating the *average* payoff each strategy receives when played against the population at large The game-theoretic description's explanatory advantages become still more evident when we compare the two accounts' competing characterizations of the further condition required in order for the cooperative strategy to evolve here for the game theorist, cooperators must be more likely to interact with one another (whether because of population structure or viscosity, the association of kin, communication, the recognition of like-minded souls, or whatever), while Sober and Wilson must represent this same condition as an increase of variation in the composition of their ephemeral "groups "

Thus, even if we accept Sober and Wilson's suggestion of theoretical equivalence between multilevel selection theory and its apparent competitors (and are consequently reluctant to deny *categorically* that these cases could be regarded as instances of something plausibly called "group selection," even if it is not quite the classical group selection attacked by Williams, Maynard Smith, and others) the central issue becomes how often it will be *useful* to think of this single process as a form of group selection Despite repeated assurances, Sober and Wilson provide few concrete reasons to think that multilevel selection theory will provide a broadly perspicuous or helpful perspective, even (as the example of two-player prisoner's dilemma with correlated interaction illustrates) in all those cases in which population structure plays an important evolutionary role

Fortunately, these considerations do not undermine the interest of the remainder of the book, for *Unto Others* goes on to place human beings centerstage, and Sober and Wilson go on to argue that ancestral populations of human beings did satisfy the special requirements of traditional group selection, amplified by the fact that our cognitive sophistication enabled altruists to selectively interact more easily, made it possible to evolve mildly altruistic systems of rewards and sanctions that enforced more dramatically altruistic behaviors, and rendered competing human groups increasingly distinct and internally homogeneous through the influence of cultural selection Sober and Wilson's anthropological evidence (drawn from a random sample of anthropologists' field descriptions of human cultures in the Human Relations Area File) ground arguments that are among the most speculative in the book, but they are also some of the most interesting, and they make at least a prima facie case deserving serious consideration (nor do the authors claim more) for the importance of group selection and the ubiquity of altruism in human evolutionary history

The second half of *Unto Others* turns to questions of psychological (as opposed to evolutionary) altruism, contrasting the case for psychological egoism (the claim that our ultimate psychological motives are uniformly self-directed) with a pluralism that includes ultimate (rather than merely instrumental) motives that are genuinely altruistic (that is, directed toward promoting the welfare of others) They effectively demolish weak philosophical and empirical arguments on each side (including misguided appeals to parsimony) and seek to break what they regard as the ensuing stalemate with evolutionary considerations using parental care as their example, they argue that considerations of reliability (in advancing fitness), availability (in ancestral conditions), and efficiency (in energetic cost) jointly favor the evolution of a motivational pluralism, including genuinely altruistic motivations over the egoistic alternatives

Sober and Wilson's evolutionary argument errs, however, in treating hedonism as the sole representative of psychological egoism This is defended by pointing out that versions of egoism which appeal to desires for external goods invariably wind up invoking hedonistic motives to explain away apparently altruistic actions, but this simply does not support the authors' conclusion that an *equally* exclusive commitment to hedonistic motives must be the most difficult form of egoism to refute[1] Indeed, the toughest opponent for the thesis of genuine psychological altruism, especially in an evolutionary context, is a *pluralistic* egoism in which our self-oriented impulses include a wide variety of internal *and* external cues that are (or were in ancestral conditions) relevantly connected to our fitness This way of thinking about human motivation is familiar from the literature of sociobiology and evolutionary psychology, and it is especially important for the authors to consider, given that the core of their argument for motivational pluralism is that a genuinely altruistic psychological mechanism for ensuring parental care will enjoy greater reliability *in promoting fitness* than the hedonistic alternative they consider

Setting aside their choice of target, Sober and Wilson do not take seriously enough the question of whether genuinely altruistic psychological motives would be easier or more likely to evolve than hedonistic counterparts After all, parental care is neither the only nor even the most important behavior that well-adapted organisms need to perform, so any organism that engages in it will need a mechanism for *trading off* its impulses toward parental care with those directed toward other important behaviors While Sober and Wilson argue (in connection with comparative efficiency) that even a hedonist requires some mechanism for trading off the various *kinds* of pains and pleasures that make up her motivational structure (including the propositionally-mediated variety hedonists will need even to treat other-directed desires instrumentally), it remains reasonable to think it more likely for an organism to evolve motivations for an important new behavior that fits into the (comparatively) common currency of an existing hedonistic mechanism for trading off each of its competing motivations with one another (the most likely ancestral condition, as Sober and Wilson recognize at least for the sake of argument) than to evolve an entirely new mechanism for trading off a new fundamental motivation against each of them

Ultimately, this pattern is representative of *Unto Others* as a whole Sober and Wilson offer a trail-blazing blend of insight, speculation, evidence, argument, and conjecture which will surely fuel serious thought and discussion in a wide variety of fields for a long time to come, even as the raw ambition of their case repeatedly leaves its supporting argumentation lagging behind, gasping and racing to catch up

P KYLE STANFORD

University of California/Irvine

The Bounds of Agency An Essay in Revisionary Metaphysics CAROL ROVANE Princeton Princeton University Press, 1998 ix + 260 p Cloth $45 00

The aim of Carol Rovane's *The Bounds of Agency* is to substitute an adequate and coherent conception of personhood for the confused mixture of animalism and Lockean psychologism which our intuitions encompass The assumption governing this effort is that despite the lack of agreement over what follows if an entity is discovered, or deemed to be a person, the possession or acquisition of the property of personhood is valuable Although it cannot be better for objects in general to be persons rather than things, it is better for human beings to be persons than nonpersons Criteria for personhood have to support the intuition of superior value, at the same time, it is risky to load the notion with too many duties or privileges held to be knowable a priori and necessary Rovane occupies the middle ground by claiming that "person" is a normative notion, though it is neutral with respect to specific theories of obligation Unlike P F Strawson's "descriptive" metaphysics of persons which reassured readers that

0022-362X/01/9801/47-54

their ordinary reactive attitudes stood in no need of further justification, Rovane's revisionary metaphysics is intended to unnerve. Our animalist prejudices, she thinks, interfere with our ability to recognize persons where they exist. Our simultaneously held psychologistic prejudices interfere with our ability to perceive social injustice because they lead us to focus on the wrong attributes as constitutive of personhood. In the end, the tension between the book's commitment to ethical neutrality and its revisionary aspirations remains unresolved. Nevertheless, there is much to learn from and to ponder in this thoughtful and deeply intelligent study.

Rovane begins by arguing that Locke was right to insist that animal identity—the organic integrity and persistence of a human body— is neither necessary nor sufficient for personal identity. But our reaction to thought experiments involving multiple realizations of copies of our mental contents or losses of large chunks of them indicates that we are unprepared for a complete break with animalism. Rather than try to weld together parts of animalism and psychologism, or give up the analytical effort, Rovane goes down a fourth, distinct path. Although she gives some attention to neo-Lockean thought experiments, she is little concerned with the problems of reidentification or with inventing or solving puzzle cases. She argues that to be a person is to be capable of engaging in "agency-regarding relations". "[T]he class of persons is the class of agents who mutually recognize one another as facing, in all of their relations, the same ethically significant choice concerning whether and when to engage in agency-regarding relations" (49). An agency-regarding relation arises between agents "when one agent attempts to influence another, and yet aims not to hinder its agency." A number of conditions must be realized for such relations to be possible. First, there must be agents who attempt with some success to carry out complex activities and pursue distant goals, and who communicate with one another in order to do so. Second, these agents must be committed to achieving what Rovane calls "overall rational unity." A creature with this commitment strives to reduce conflicts and inconsistencies in its belief sets, to order its preferences in noncontradictory ways, and to act in ways consistent with its beliefs. Third, these agents must be able to recognize other agents like themselves, and must be recognizable by those other agents.

Persons reveal their commitment to overall rational unity in internal deliberation and self-criticism, but also through consultation with one another. Agents give one another advice, discuss and criticize each other's plans, point out limitations and possibilities that have been overlooked and implications that were unseen, and they are naturally receptive to taking advice. But they also threaten, manipulate, deceive, mislead, and even mesmerize each other, and they are also unfortunately receptive to these efforts to direct, redirect, and hinder their agency. Generally speaking (though some minor qualifications are recognized on pages 94-95), to the extent that one's relations with some entity consist of efforts to influence its behavior by lies, threats, or the provision of incentives, one is not treating it as a person; to the extent that some creature cannot be influenced except in this way, it is not a person. Rational influence, by contrast, "always proceeds by trying to find and present reasons that another person will find compelling in light of its own rational point of view" (137). Psychologism has some truth, insofar as the persistence of memory, the capacity to project oneself into the future, and relative stability of preferences and dispositions, help to diagnose personhood. But "persons endure over time only insofar as they have commitments to unifying projects, and these unifying projects give them renewed reasons, at the different moments of their lives, to strive for overall rational unity over time" (165). No person can introspect its own personhood, except by discovering its capacity to exercise agency and its commitment to rational unity, and no external observer can individuate persons by individuating human bodies, not even fully-grown, wide-awake, bodies that have not been subjected to comissurotomies, transplantations, implantations, graftings, splittings, and similar procedures. We choose to engage in or refrain from entering into agency-regarding relations, and in that sense we decide whether certain other human animals are persons, or fully persons, or persons all the time, and we also decide whether to be persons vis-à-vis particular others.

Rovane denies that any specific obligations toward human beings or persons follow from her characterization. The moral program of an egoist or a utilitarian might require him to engage in agency-regarding relations with no one, or with very few actual persons in order to realize the goal of bringing about the greatest happiness for himself, or for the greatest number. At the same time, she claims that nonacknowledgment or partial acknowledgment of personhood was once common, and she implies that such patterns may be residual even in cultures that enjoy an enlightened discourse of rights and respect. Some humans may engage some other humans effortlessly in rational discussion while obstructing these others or failing to assist them in formulating their own long-term plans. They may attempt to interfere with their agency by threats, bribes, and mesmerization. And they may convince themselves that dealing with them otherwise is impossible.

> Sometimes, agents wish to retreat from engagement with other agents as person, but without facing up to the fact that they have thereby made and carried out an ethically significant choice They do this by denying that the agents who occasion the choice are really persons By denying this, they manage to deny that these agents have any preferences in favor of being engaged with as persons Indeed they manage to deny that such engagement is even possible (119)

Rovane characterizes the attitude toward semipersons as "hypocritical " If legislators and officials in earlier centuries had believed without reservation that women could not think well, manage money, or perform surgical operations, there would have been no reason to pass new laws prohibiting the holding of certain offices and participation in certain activities or to defend the old laws, they both acknowledged and did not that women were persons Rovane does not embrace the claim that differential restrictions on humans violate universality requirements and are always wrong, the hypocritical posture, she thinks, might be a psychological precondition of furthering some moral aim (121) It is only wrong, she thinks, to deny that the hypocritical posture is of ethical moment (6)

Despite the appeal to the higher mental faculties, Kantians will not like this We are told that the commitment to overall rational unity consists in

> resolving contradictions and other conflicts amongst one's beliefs and other psychological attitudes, accepting the implications (both logical and evidential) of one's attitudes, ranking one's preferences, identifying opportunities for action, assessing the probable consequences of performing the actions that are open to one, determining what means are available for achieving one's ends, evaluating one's ends, reassessing the relative preferability of one's ends in light of both the available means to achieving them and the probable consequences of acting on them (139)

Rovane concedes that not all human animals engage in this optional mental work, and she emphasizes that there is no categorical requirement for them to do so In some respects, the admissions qualifications to personhood are set low A person need strive only to acquire well-ordered beliefs and consistent preferences to earn minimal ethical consideration, not true beliefs or good preferences As a result, the reason for the normative carryover from personhood to even minimal entitlement remains unclear On the other hand (though some effort is made to avoid overintellectualizmg (146-47)), the requirements are set high Because Rovane is interested in the plight of persons-who-are-not-recognized to be such, she represents it as unconscionable that there should be persons who are fully committed to rational unity who still receive none of the privileges of personhood But most human beings are acting, living, and preserving their beings, as Baruch Spinoza says, but not devoting much effort to uncovering and reconciling latent contradictions in their belief systems, and most do not have careers or life-long ambitions either Suppose a creature brought up in constrained circumstances—in a harem, for example, or in a family of slaves, fails to exert and to signal its agency, it devotes no effort to cognitive maintenance and upgrades and forms no long-term plans It is not on Rovane's criteria a person and has no preference for being regarded as such, therefore no harm is done when the agents around it refuse to engage it in agency-regarding relations Most Kantians think there is something wrong with this Now it might be a mere prejudice that most members of the species homo sapiens had better come up persons on any account Why should "person" not be a somewhat exclusive designation, reserved for a few? But the problem is that Rovane's conditions for personhood are in most respects hard, while the entitlements they generate are soft Being a person does not carry with it an automatic right to noninterference or regard, but only to the minimal ethical consideration involved in someone's having to decide that it is all right to interfere with you or dismiss you, and having the mere potential to be a person, being potentially educable, potentially project oriented, and so on seems to carry no entitlement at all So what the theory really says in the end is that, if you are one of those creatures which has a commitment to rational unity and long-term plans, you ought to be treated as such unless there are overriding moral reasons not to But this is not a *metaphysical* theory of personhood, it is a theory concerning a widespread form of social injustice

This argument up to here occupies roughly the first half of the book The second half is devoted to a further attack on animalism and psychologism through a demonstration of the possibility of multiple persons within a single human body, and group persons that lack the introspective capacities of a typical Cartesian-Lockean consciousness Although corporate persons have been discussed extensively in the business-ethics literature, and although multiple personalities have come into prominence with Ian Hacking's skeptical treatment, there are several innovations here One is a redeployment of Thomas Nagel's technique of modeling altruistic behavior on prudential behavior in *The Possibility of Altruism*[1] to meet the objection that group persons lack the introspective abilities needed to pursue intentional goals Group persons do not require introspective access to shared memories or other subjective accoutrements, they need have only a commitment to overall rational unity with respect to some set of future-directed plans The second is a redeployment of H P Grice's theory of communication In prudential behavior we anticipate communicating with our past selves by recognizing their previous intention to affect us in a certain way, and group persons develop analogous projective relations with their future selves (149ff)

---

[1] New York Oxford, 1970

The possibility that more than one full person can inhabit a single body is given content through a sympathetic and credible recounting of the case of Sybil, the famous "multiple," but also through the exploration of the idea that we individual human-sized persons are more like Sybil than we might at first think We typically have commitments to many distinct long-term activities, each of which requires a sustained coordinated effort over time, and each of which makes different demands on our cognitive ordering and forecasting abilities (174) Working as an academic, making progress with some musical instrument, and keeping a household going are essentially unconnected and even conflicting projects, and somewhat different personalities are required for party going and for chairing meetings Unlike Sybil, most of us try to make a single human life out of these separate streams of activity, and personality-drag, which Sybil did not experience, ensures that some of our underlying frivolity or dourness gets ported from one venue to another With a little imagination, it is easy to see that our sublives could take on the isolation of Locke's Day-person and his Night-person

Although the endorsements on the back cover suggests that *The Bounds of Agency* is a convincing or at least very serious defense of the possibility of group and multiple persons, the book's intention is not I think to provide such a defense Rovane insists that she is interested only in the pure possibility of nonstandard persons and takes no position on whether we actually meet with them in space and have to deal with them appropriately Very little is said about how to recognize a group or multiple person and what to do if faced with one Rovane's principle of moral neutrality implies that, if two persons inhabited a single body, it might be morally permissible to deny both their agency by refusing to allow them to marry two spouses or collect two government pensions, even if their affiliation and consumption preferences and associated plans were wholly irreconcilable But this is tantamount to conceding that animalist and psychologistic beliefs do enter into our notion of personhood and can overrule the alleged ethical criterion And Rovane seems to be less interested in setting out a distinct notion of the pure possibility of a nonstandard person than in showing that we meet with indistinct versions of nonstandard persons routinely and in space The book's message—contrary to what its title suggests—is that our agency is *not* bounded but is constituted by our identification with programs we selectively take up and cast aside, and with our being seen as doing so and as capable of doing so by others "The ethical criterion [of personhood] suggests that these boundaries [between individual human animals] do not need transcending, because we already live on both sides of them and moreover we can always redraw them" (208)

As the author points out, we are surrounded by and participate in numerous joint enterprises—intellectual partnerships, social and political organizations, marriages These enterprises can bind people into units which have their own interests and which must perform the kind of planning and cognitive maintenance that, according to Rovane, all real persons do spontaneously Under ideal conditions, nonstandard persons keep track of one another's contributions and thoughts about their project, have shared beliefs including beliefs about others' commitment to the project, and attempt to influence each other rationally about the direction of the project (145)

This is a well-developed, complex, and gracefully written book, it is brimming with ideas It positions itself in the largely unoccupied space between the morally serious but philosophically feeble literature on the social construction of the self and the rigorous but just-for-fun literature of splits and fusions, it is rigorous and serious, too It invites a reconsideration of the connection between voluntary and structural curtailments of agency, and respect for persons, though it deliberately downplays its substantive moral message Philosophers of science discovered the interest of the problems of collective rationality and irrationality and have been exploring them since, but standard individual projects continue to be focal in moral philosophy *The Bounds of Agency* suggests that it will prove equally rewarding to look at the ethical features of small-scale cooperative enterprises

Not only Kantians but also animalists will have their reservations nevertheless A solid treatise by an animalist on personhood has yet to be written, but I think it would say something like this The significant properties of human beings, including their capacity and tendency to think of themselves as persons entitled to moral consideration, as well as their capacity and tendency to regard other humans sometimes, but not always as persons, are explained by facts about the kinds of animals we are and the instructions encoded within our specific type of animal bodies The intuition of a self-same self has a biological function, and the ability to reidentify other human animals as the very same ones, and to ascribe to them self-same selves despite changes of appearance and context is important for us Other facts about our animal bodies and their histories explain why interference and manipulation and the refusal to engage in agency-regarding relations are ubiquitous For the animalist, our theoretical interest in the metaphysical question of personal identity reflects the elevation of a practical problem, critical to our lives as social animals, into a theoretical problem calling forth our native ingenuity, our skills at the rotation of figures, the permutation of possibilities, the detection of isomorphisms, these transformative experiments being typical mental amusements of our species Rovane's argument against animalism is the intuition that several persons can exist in one body and that a number of human bodies can be associated with one person But all proposed criteria for personhood, even the most rationalistic, depend on our endowments as long-lived creatures that happen to live by our wits, our hands, and our capacity for speech, as the other animals do not Although group and multiple persons are theoretically possible, the theory of persons does not explain the empirically observed fragility of groups that are committed to a shared rational unity and the psychological difficulty of living a multiple life We humans can defect from any group just by walking away on our animal feet, and we often do, but we are stuck with ourselves, and even if doing so would lead to certain efficiencies and minimize conflicts we cannot fragment into more persons and personalities than our physical constitutions allow

CATHERINE WILSON

University of British Columbia

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO 2, FEBRUARY 2001

THE PRINCIPLE OF SUFFICIENT REASON

Many attempts to refute Gottfried Leibniz's *Principle of the Identity of Indiscernibles* (PII) and *Principle of Sufficient Reason* (PSR) turn upon symmetries If a symmetric world were possible, then PII would stand refuted, since such a world would include distinct objects sharing all of their qualitative and relational properties Similarly, if there were a class of possible worlds whose shared geometry and laws were invariant under some set of symmetries, then PSR would stand refuted, since there would exist distinct worlds sharing all of their qualitative and relational properties think here of the standard approach to Newtonian mechanics, under which the symmetry of the theory under Euclidean isometries implies the existence of distinct worlds with qualitatively identical geometries and arrangements of matter, differing only as to where the matter is placed in space

Ian Hacking[1] once observed that symmetry-driven counterexamples to PII are inherently inconclusive One can always deny that the proffered description manages to characterize any possible world accurately, maintaining instead that the only genuine possibility in the neighborhood arises when the qualitatively identical objects of the original description are taken to be numerically identical

I claim that PSR can be given a similar line of defense a description of a set of possible worlds which includes pairs of worlds with identical qualitative structures can always be taken to correspond to the sparser set of possibilities which arises when qualitatively identical worlds are identified

My argument has two parts, held together by an admittedly loose interpretation of Leibniz the later sections form a commentary on

---

[1] "The Identity of Indiscernibles," this JOURNAL, LXXII, 9 (May 8, 1975) 249-56

0022-362X/01/9802/55–74

some theorems concerning symmetries and the structure of the spaces of physical possibilities of classical mechanics, the first few sections are intended to set this discussion in context.

I. HACKING'S STRATEGY

Consider Max Black's[2] famous challenge "Isn't it logically possible that the universe should have contained nothing but two exactly similar spheres" (*ibid.*, p. 156)? If this is granted, then "every quality and relational characteristic of the one would also be a property of the other" (*ibid.*), and PII is soundly refuted.

Symmetry plays a crucial role in arguments like Black's. A two-sphere world is a potential counterexample to PII only if the spheres are images of one another under reflection in the plane bisecting the segment joining their centers—if they were different sizes or colors, for instance, PII would of course be perfectly safe. Similarly, if we describe a two-dimensional Euclidean world containing three massive point-objects, then PII is threatened if and only if there is a symmetry of space which maps the triangle determined by the three points onto itself while preserving the masses of the vertices. Thus, no scalene triangle threatens the principle, while an isosceles triangle poses a threat if and only if reflection in its axis of symmetry interchanges vertices of the same mass. An equilateral triangle poses a threat if and only if some pair of vertices are the same mass (in which case masses are preserved by reflection in the axis perpendicular to the segment that joins these points), if all three vertices are the same mass, then masses are also preserved by rotations of 120° and 240° degrees about the center of the triangle (these rotations permute the vertices without interchanging any pair).

More generally, suppose that we are given a description of a set of material bodies, possessing certain properties and standing in certain (Euclidean) geometric relations to one another. If this arrangement is symmetrical—so that there is some nontrivial combination of reflections, rotations, and translations which maps the arrangement onto itself in a way that preserves properties—then, as in Black's example, there will be a pair of objects that share all of their qualitative and relational properties in virtue of playing identical roles in the pattern of relations and properties instantiated. Thus the situation described, if a genuine possibility, constitutes a counterexample to PII. On the other hand, if there is no such symmetry of the situation, then PII is in effect, since each object is distinguished from every other in virtue of some property it possesses, or the role that it plays in the pattern of geometric relations holding between the objects

---

[2] "The Identity of Indiscernibles," *Mind*, LXI (1952) 153-64.

Hacking's defense of PII exploits the close relationship between PII and asymmetry His paper proceeds via ingenious elaborations of a simple theme any description of a symmetric arrangement of bodies can be taken to be a *mis*description of an asymmetric state of affairs involving fewer bodies When Black asks Leibniz to imagine a world containing only two identical iron spheres, Leibniz can respond—as he did to Samuel Clarke—that "To suppose two things indiscernible, is to suppose the same thing under two names "[3]

When presented with a story about a world comprising only two identical drops of water whose surfaces reflect one another, Hacking's Leibniz responds that what has in fact been described is a world containing a single drop whose surface reflects itself Intuitively, this one-drop world is constructed from the 'world' originally described by identifying points of space and bits of matter which are mirror images of one another

Opponents of PII can offer more and more elaborate worlds as counterexamples But because these worlds will always involve symmetric arrangements of matter, proponents of PII will be able to defuse the threat by identifying qualitatively identical objects to generate asymmetric worlds that obey the principle Implementing this strategy may sometimes require unseemly contortions Some truths about the proffered symmetric world will carry over to its asymmetric replacement ('Each drop reflects some drop'), some truths will not ('No drop reflects itself') Thus, as Hacking notes, it can happen that the symmetric world admits a simple description, beside which any description of the asymmetric world appears baroque Hacking insists, though, that this sort of thing can at most embarrass Leibniz—it cannot force him to abandon PII

II ONCE MORE, ABSTRACTLY

I think it helpful to have in view a more abstract description of Hacking's strategy (Readers who do not like the sound of that may prefer to skip this section) We are interested in structures and their symmetries [4] A *structure* consists of a set of objects, *A*, together with a set, $\{R\}$, of relations on A (including, possibly, one-place relations—namely, properties) [5] A symmetry of a structure is a permutation of its objects which fixes each of its relations [6] We equip ourselves with a first-order language with identity which includes a predicate symbol for each relation of our structure, a set of variables that range over objects, and a set of constants large enough to provide a name for each object [7]

---

[3] See H G Alexander, ed , *The Leibniz-Clarke Correspondence* (New York Barnes and Noble, 1956), L IV 6 Here 'L IV 6' indicates the sixth paragraph of Leibniz's fourth letter Future references to this volume will be noted in the text using this same abbreviation

[4] What follows is a sketch of some elementary model theory, phrased somewhat idiosyncratically This material can be found in, for example, Wilfred Hodges, *A Shorter Model Theory* (New York Cambridge, 1997), see his §§1 1–1 4, 2 3, and 8 3,

This language provides two principal modes for describing our structure We can use variables but no names, and focus on *theories*—(arbitrary) sets of true sentences about our structure which employ neither constants nor free variables Alternatively, we can choose to employ names but no variables setting up a *nomenclature*—an assignment of a constant of our language to each object of the structure—we can regard the constants as names for objects, and formulate true statements about the objects named In particular, each nomenclature is associated with a *complete description*—the set of all true atomic sentences about the structure relative to this system of naming A complete description of a structure does what no theory about an infinite structure can do—determines it up to isomorphism

Let us fix a structure, and consider a nomenclature and its associated complete description Each permutation of the set of objects of our structure induces a new nomenclature and a new complete description Symmetries are the permutations that leave invariant our initial complete description Thinking of symmetries as permutations of names rather than objects, we can see that they leave invariant the list of true sentences describing the structure

The existence of a symmetry mapping object a to object b indicates that a and b play the same role in our structure Thus, if a and b are distinct objects in a structure describing a possible world then we have a counterexample to PII Conversely, if a structure admits no

---

keeping in mind that the map that we are interested in below—under which a is sent to $[a]$—is a subjective homomorphism which preserves one-place relations

[5] It is sometimes convenient also to include in our structures functions on their sets of objects Taking these into account here would be straightforward enough, but unwieldy—so I proceed with relations alone for the time being, but help myself to functions later

[6] A permutation is a bijection from the set of objects to itself Any such map also acts on n-tuples of objects in an obvious way (that is, entry-wise), and so maps each n-place relation—conceived of as a set of n-tuples, on which the permutation acts element-wise—to an n-place relation The upshot is that a permutation ϕ is a symmetry if $R(a_1,\ ,a_n)$ if and only if $R(\phi(a_1),\ ,\phi(a_n))$

[7] This could be a very large language—no restriction is placed on the cardinalities of the sets of objects and relations

symmetries, this indicates that each of its objects plays a distinct role[8] Thus a structure (of physical objects) contains nonidentical indiscernibles if and only if it admits a nontrivial symmetry

We can now describe Hacking's strategy as follows Having fixed an arbitrary structure, we construct a closely related *quotient structure* by identifying objects related by symmetries We define an equivalence relation on the set of objects of our original structure by declaring that $a{\sim}b$ if there exists a symmetry of that structure, $\phi\ A\rightarrow A$, with $\phi(a)=b$, we write $[a]$ for the equivalence class of a, $\{b\ \ a{\sim}b\}$ The set of objects of the quotient structure is $\{[a]\ \ a\in A\}$, the set of equivalence classes of objects of the original structure The quotient structure has a relation $[R]$ for each relation R of the original structure, $[R]$ is the smallest relation such that if $R(a_1,\ \ ,a_n)$ for some n-tuple of objects of the original structure, then $[R]([a_1],\ \ ,[a_n])$

We can use the same language to describe both structures (using the same predicate symbol, '$\mathfrak{R}$', as a name for both R and $[R]$) Choosing names for our new objects generates a complete description of the quotient structure This will coincide with a complete description of the original structure (relative to some nomenclature) if and only if the two structures are isomorphic—that is, if and only if the original structure possesses no nontrivial symmetries

Even when the original structure does admit nontrivial symmetries, the complete descriptions of the two structures will be closely related A complete description of the original structure can be transformed into a complete description of the quotient structure by taking names of objects related by symmetries to name identical objects[9] Thus, a complete description of the original structure can be taken to be the sort of misleading description of the quotient structure which results when some objects are given multiple names

We can also consider the relations between theories describing the two structures Some sorts of (constant-free) sentences are true of the quotient structure if they are true of the original structure This holds for sentences which are negation-free and those which mention only one-place relations. More can be proved: if a sentence is in negation normal form (so that all of its negation symbols apply to atomic formulas) and all of its negation symbols apply to one-place relations, then it is true in the quotient if it is true in the original structure.[10] But not too much more: suppose that the original structure includes an irreflexive binary relation, R, and objects a and b such that $R(a,b)$ and $a\sim b$, then in the quotient structure, we have $[R]([a],[b])$, with $[a]=[b]$, so the sentence '$\forall x \neg\mathfrak{R}(x,x)$', attributing irreflexivity to R and $[R]$, is true in the original structure but false in the quotient structure.

---

[8] Note the distinction between the symmetries under discussion here and the sort of permutations that drive Hilary Putnam's famous argument about reference We are interested in the nontrivial *automorphisms* of a structure—nontrivial permutations that *fix* each of the relations of a given structure There are many structures—such as well-ordered sets—which do not admit any nontrivial symmetries Putnam, on the other hand, shows that any (reasonably rich) structure is isomorphic to a structure which has the same set of objects but whose relations are generated by arbitrary permutations (and which, therefore, will typically be distinct from the relations of the original structure), see the Appendix to *Reason, Truth, and History* (New York Cambridge, 1981)

[9] This would not be true if we had included nonatomic statements in our complete description—consider the fate of a statement like '$a\neq b$' when $a{\sim}b$

The upshot: whenever we have a structure that admits nontrivial symmetries, we can factor these out, constructing a quotient structure closely related to the original one. In special cases—such as when all of the relations of the structures are one-place—every theory true of the original structure is also true of the quotient.[11]

But in general the two structures will be described by related but distinct theories. Hacking argues that the committed advocate of PII can brush off one worry aroused by this divergence: that theories describing the quotient world may be complex compared to those describing the original structure.

There is, however, a second worry in the neighborhood: theories of the quotient may be too *simple* in comparison to those of the original world. The following cases were suggested to me by Kit Fine. Consider two structures for a given countable set of objects, in one structure the set of objects is given an ordering isomorphic to the integers, in the other an order isomorphic to the rationals. The quotient of each structure is just a single object, related to itself. Here we have somewhat rich structures that collapse to a disappointingly simple one when we pass to the quotients. How sure can we be that taking the quotient of a 'world' will yield a possible world?

Again, the committed Leibnizean need not be troubled: "Things which are uniform, containing no variety, are always mere abstractions."[12] But the more moderate among us will probably conclude that both of these worries render it necessary to examine the relation between a structure and its quotient on a case-by-case basis in order to discern whether the two can indeed be taken to correspond to the same sorts of complexes of possibilities

---

[10] For a deeper result that goes in the opposite direction, see Hodges, *Shorter Model Theory*, Theorem 8.3.3(b).

[11] This special case assumes premier importance for those who, following Bertrand Russell's *A Critical Exposition of the Philosophy of Leibniz* (New York: Routledge, 1992), have Leibniz maintaining that all facts can be reduced to properties of monads. For a recent defense of this line, see J. A. Cover and John O'Leary-Hawthorne, *Substance and Individuation in Leibniz* (New York: Cambridge, 1999), chapter 2.

[12] *New Essays on Human Understanding* (New York: Cambridge, 1996), §II.i.2.

III HACKING'S CONCLUSION

Hacking sets out to show that PII can be defended against a certain sort of counterexample presentations of symmetric worlds He proposes that any description of a symmetric world, in which multiple objects play the same role in the pattern of properties and relations instantiated, can always be taken to be a misdescription of an asymmetric world that respects the principle That this strategy is available—even if there is room to doubt its prudence in many cases—shows that PII can be protected from counterexample by spatiotemporal symmetry If we indulge in a little rational reconstruction, and take PII to amount to the denial of the possibility of symmetric worlds, then we reach Hacking's conclusion—"the pursuit of logical questions might settle the issue of [PII], but mere reflection on spatiotemporal examples is never enough" (*op cit*, p 249) [13]

Hacking[14] himself has proposed a conception of the nature of logic which does imply a version of PII At the same time, it fails to underwrite the catechetical principle that identity implies necessary identity [15] In defending this lapse of orthodoxy, Hacking registers his doubts about other views concerning rigid designation which form the very basis of the faith He rejects the claim that in naming two actual qualitatively identical drops of water *Jake* and *Jesse* we gain the ability to describe two distinct possible worlds *Beta* and *Gam*, the first inhabited only by Jake, the second only by Jesse

> by cold naked stipulation we produce not two possibilities but one Beta and Gam, for all our naming, are the same world, populated by one drop, with just so many properties I am not here arguing a case for or against "trans-world identification " I am arguing that there is no sense in postulating two distinct worlds which have no point of distinction except in the sheer names of their inhabitants (*ibid*, p 626)

---

[13] Leibniz, however, seems to have had a much broader conception of the content of PII at L III 5 he appears to use it to rule out the existence of pairs of qualitatively identical worlds, it also appears to forbid the existence of perfectly similar leaves, drops of water, or eggs in our own asymmetric world (for example, at L IV 4 and L V 2, for further references and discussion, see §VII 3 of Benson Mates, *The Philosophy of Leibniz Metaphysics and Language* (New York Oxford, 1986))

[14] "On the Reality of Existence and Identity," *Canadian Journal of Philosophy*, VIII (1978) 613-32, here pp 618-19 See also his "What Is Logic?," this JOURNAL, LXXVI, 6 (June 1979) 285-319

[15] "On the Reality of Existence and Identity," pp 620-21

As Hacking goes on to point out, "this is not an argument about the identity of indiscernibles," as he construes that principle [16]

But the question of the existence of such pairs of qualitatively identical worlds is intimately involved with PSR, as Leibniz deployed it in his dispute with Clarke

IV LEIBNIZ AND CLARKE

Early in his correspondence with Clarke, Leibniz asserts that the reversal of the decay of natural religion in England can never be accomplished by means of that country's native, mathematical, principles alone—metaphysical principles are required The principle of contradiction suffices to found arithmetic and geometry, but "in order to proceed from mathematics to natural philosophy, another principle is requisite, as I have observed in my *Theodicy* I mean, the Principle of Sufficient Reason, viz that nothing happens without a reason why it should be so, rather than otherwise" (L II 1)

Archimedes's[17] postulate that "Equal weights at equal distances are in equilibrium, and equal weights at unequal distances are not in equilibrium but incline towards the weight which is at the greater distance" is a special case of this principle Much of the controversy between Leibniz and Clarke hinges on the question of the extent to which God—unlike a balance or an ass—is capable of *arbitrarily* choosing between qualitatively identical alternatives In pressing this question, Clarke raises the following difficulty

> why this particular system of matter, should be created in one particular place, and that in another place, when, (all place being absolutely indifferent to all matter,) it would have been exactly the same thing *vice versa*, supposing the two systems (or the particles) of matter to be alike, there can be no other reason, but the mere will of God Which if it could in no case act without a predetermining cause, any more than a balance can move without a preponderating weight, this would tend to take away all power of choosing, and to introduce fatality (C II 1)

Leibniz never answers this objection to Clarke's satisfaction any more than Clarke ever demonstrates to Leibniz his comprehension of the great principle under discussion

But the particular example broached by Clarke in this passage leads to one of the most interesting of the many proliferating side issues under discussion Leibniz jumps at the chance to air his complaints about the Newtonian doctrine of absolute space He opens with a bold bluff "I have many demonstrations, to confute the fancy of those who take space to be a substance, or at least an absolute being But I shall only use, at the present, one demonstration, which the author here gives me occasion to insist upon" (L III 5) He goes on to recapitulate Clarke's own argument "'tis impossible there should be a reason, why God, preserving the same situations of bodies among themselves, should have placed them in space after one certain particular manner, and not otherwise, why every thing was not placed the quite contrary way, for instance by changing East into West " We are now to read this argument, however, not as a refutation of PSR by counterexample, but as a reductio of the supposition of absolute space which proceeds via an appeal to that principle This clears the way for Leibniz's own conception of space

---

[16] Again, compare L III 5

[17] The first postulate of *On the Equilibrium of Planes* in Thomas Heath, ed , *The Works of Archimedes* (New York Dover, 1953), p 189

> But if space is nothing else, but that order or relation, and is nothing at all without bodies, but the possibility of placing them, then those two states, the one such as it is now, the other supposed to be the quite contrary way, would not at all differ from one another Their difference therefore is only to be found in our chimerical supposition of the reality of space in itself But in truth the one would exactly be the same thing as the other, they being absolutely indiscernible, and consequently there is no room to enquire after a reason of the preference of the one to the other (L III 5)

Where Clarke sees many possible arrangements of bits of matter consistent with fixed specification of their relative distances, Leibniz recognizes only one We can think of Clarke as generating his space of possibilities by considering some particular way of embedding a material system into Euclidean space which gives rise to the specified relative distances, then shifting, reflecting, and rotating this material distribution in absolute space to generate many more, each satisfying the same set of relations of relative distance This is possible because of the "indifference" of space to matter, which also guarantees that each of the alternatives so generated will be qualitatively identical (in one world, *this* object is *here*, in that one, *there*, but these two parts of space share all of their qualitative properties and relations) Leibniz protects PSR by identifying alternatives related by the symmetries of Euclidean space For him there is a single possibility here the bodies, in space, with such-and-such relations between them

V PROTECTING PSR

Notice that the symmetries at stake in the Leibniz-Clarke correspondence are not symmetries of (typical) *worlds*—only very special arrangements of matter are invariant under rotations, reflections, or translations Rather, the symmetries in question are symmetries of the *space of possibilities* consider the set of possible dispositions of some material system in absolute space, act on each possibility by (say) translating each particle by some given amount in some given direction, the *set* of possibilities is invariant under this action (and under the actions of reflections and rotations as well), although few of its members are

Part of the content of PSR involves the denial of the existence of qualitatively identical worlds [18] In this guise the principle appears to be vulnerable to symmetry-driven counterexamples By analogy with Hacking's defense of his construal of PII, we see, however, that this part of the content of PSR can be protected from this class of putative counterexamples if we insist that symmetries should be factored out of the space of possible worlds

Note that the apparatus of section II can be deployed whenever we have a set structured by some relations In Hacking's defense of PII, the structured sets in question were possible worlds, whose objects corresponded to material objects, endowed with properties and standing in geometric relations to one another Hacking advises us to factor out symmetries from such sets to ensure that they do not include objects sharing all of their qualitative properties and relations In order to defend PSR against symmetry objections, we need to turn our attention to structured sets whose objects are possible worlds, factoring out symmetries from such sets to ensure that there are no distinct but qualitatively identical worlds

Mathematical treatments of classical mechanics provide one of the few settings in which structured spaces of possible worlds arise in a natural way—and also constitute the natural context for the questions raised by Leibniz and Clarke So let us focus on that case We are given a *phase space* whose points represent physically possible instantaneous dynamical states of our system—initial data Alternatively, since the physics is deterministic, we know that the space of physically possible histories must be isomorphic to the space of initial data (choosing a time at which to pose the initial data determines an isomorphism) So we can, as it suits us, take the points of a phase space to correspond to either possible worlds or to temporal slices of possible worlds

---

[18] Leibniz has uses for the principles that do not fit this mold Archimedes's principle mentioned above is one Another, very prominent, one is as a principle of maximization see the passage (from a letter to Princess Caroline) which is customarily printed as a postscript to L IV, or the letter to Des Bosses of August 19, 1715 (Leroy Loemker, ed , *Gottfried Wilhelm Leibniz Philosophical Papers and Letters* (Boston Reidel, 1970), p 613)

A classical mechanical phase space carries a sort of geometric structure that wants only the specification of a distinguished function on phase space, the *Hamiltonian*, to determine a unique curve through each of its points[19] If we think of points of phase space as representing instantaneous states, then each dynamical trajectory describes a physically possible history of states[20] In the most common case, the phase space parameterizes the set of possible positions and momenta of some set of particles or fields in Euclidean space Such a phase space carries a canonical geometric structure, and the standard dynamics arise when the Hamiltonian is taken to be the function which assigns to each state its total energy Typically our Hamiltonians are written as sums of kinetic energy terms (encoding inertial effects) and potential energy terms (encoding information about the action of forces)

Now let us suppose, as is in fact quite typical, that we have a classical mechanical theory which is invariant under some symmetries—so that there are bijections from the phase space to itself which leave invariant the geometric structure and Hamiltonian of our theory The set of trajectories through two points of phase space related by such a symmetry will have the same structure (since the dynamics are determined by the geometric structure and the Hamiltonian, which are themselves invariant under the symmetry in question) And this is to say that at the level of description we are working at—where all and only dynamical variables are included in our descriptions of possible states and worlds—that these two points represent qualitatively indistinguishable possible states or worlds So the existence of symmetries of our dynamical theory poses a threat to PSR

The advocate of PSR will observe that this threat can be deflected if we hold that it is not this original phase space but its quotient structure—the space of possibilities which arises when points of the original phase space related by symmetries are identified—which parameterizes the space of dynamically possible worlds

What is the price of this defense of PSR? We started with an elegant mathematical formulation of a physical theory In moving to a theory of equivalence classes of our original states, it is natural to worry that we will end up with something hopelessly ad hoc, perhaps with little claim to be considered a physical theory in its own right In the next few sections, I shall make some remarks about theorems in geometric mechanics which ought to allay this fear [21]

---

[19] What is needed is a means—typically provided by a *symplectic form*, a closed, nondegenerate two-form—for establishing a correspondence between vector fields and differential forms, one then solves for the vector field corresponding to dH, where H is the Hamiltonian, and integrates to find the dynamical trajectories

[20] If we think of points of phase space as corresponding to possible worlds, then two worlds lie on the same dynamical trajectory if and only if they contain the same sequence of states "shifted in absolute time "

VI EXAMPLE CELESTIAL MECHANICS

Let us apply this strategy to a familiar example—the classical theory of N gravitating point particles It is helpful to have two spaces in view here the $3N$ dimensional *configuration space* of the system (encoding the possible instantaneous dispositions of the N particles in three-dimensional Euclidean space), and the corresponding $6N$ dimensional phase space (parameterizing the possible instantaneous positions and velocities of the particles) [22] The phase space carries a canonical geometric structure, and the Newtonian dynamics are determined by taking as our Hamiltonian the sum of the kinetic energy and the gravitational potential energy

If we consider some isometry of Euclidean space—a product of translations, rotations, and reflections—then we can see how it maps each point of the configuration space of the theory to some other allowing the isometry to act on each of the positions of the individual particles leads to a new configuration of the system Of course, since isometries preserve distance, the initial configuration and the transformed configuration instantiate the same pattern of relative distance relations between particles

This action of the Euclidean group on the configuration space induces an action of the group on the corresponding phase space If we know how a given Euclidean isometry acts on each point along a curve in the configuration space, then we know how it acts on the curve itself, and this allows us to define an action of the group on tangents to curves in configuration space, which is just to say that we know how each isometry transforms the velocities of each of our particles The action is, in fact, straightforward a translation does not alter a velocity vector, a rotation or reflection acts on such a vector in the obvious way (rotating or reflecting it)

Now the important point for us is that the group of Euclidean isometries is, in fact, a symmetry group for our theory, leaving invariant the Hamiltonian and the geometric structure of the phase space—and, therefore, mapping dynamical trajectories to dynamical trajectories.[23]

---

[21] The best short introduction to the body of work discussed below is Jerrold Marsden, "Steve Smale and Geometric Mechanics," in M Hirsch, Marsden, and M Shub, eds , *From Topology to Computation Proceedings of the Smalefest* (New York Springer, 1993), pp 499-516 For more comprehensive treatments, see Marsden's *Lecture on Mechanics* (New York Cambridge, 1992), or Victor Guillemin and Shlomo Sternberg, *Symplectic Techniques in Physics* (New York Cambridge, 1984)

[22] It is harmless here and below to speak of 'velocities' where 'momenta' would be more proper

The relationalism about space manifested in Leibniz's replies to Clarke suggests the following line of thought.[24] The standard configuration space overcounts possible configurations of particles by recognizing many possible ways to instantiate each geometrically possible pattern of relative distances. We ought to replace this configuration space by the *relative configuration space*—parameterized by the geometrically possible patterns of relative distances between the particles—which arises when we identify points in the standard configuration space related by the action of the Euclidean group. We ought then to formulate a *strictly relational* dynamics, set in the *relative phase space* (parameterized by the possible instantaneous relative distances and velocities for the particles).

It is easy to see, however, that this relative phase space is not rich enough to capture the empirical content of Newton's theory: if I merely specify the initial relative distances and velocities of two (or more) particles, I have not given enough information to fix the future evolution of their relative distances and velocities—depending on the total angular momentum of the system, the particles may eventually fall toward one another, or they may form a stable or expanding orbit.

So a strictly relational reformulation of Newton's theory is not to be had. What bearing does this fact have on our general strategy for protecting PSR? Let us count dimensions. The standard phase space has a $3N$ dimensional configuration space and a $6N$ dimensional phase space, and is invariant under the six dimensional group of Euclidean symmetries. It follows that the relative configuration space is $3N-6$ dimensional and the relative phase space (which includes the relative configuration variables and their rates of change) is $6N-12$ dimensional.

We can also construct a *reduced phase space*, by taking the quotient of the standard phase space by the action of the Euclidean group

---

[23] The canonical (cotangent bundle) symplectic structure is invariant under any transformation which is the lift of a diffeomorphism of the configuration space; the kinetic energy is Euclidean, deriving from the Euclidean structure of the space in which the particles move; the gravitational potential energy sees only relative distance between particles.

[24] This program can also be motivated by relationalism about motion; see my "Geometry and Motion," forthcoming in *The British Journal for the Philosophy of Science*. Leibniz sometimes affirms the relativity of all motion; in the same texts, however, he also recognizes a notion of "absolute true motion," to be founded upon a metaphysical notion of force. Compare L V 52 and L V 53; see also Part II of "Specimen Dynamicum" in Loemker, ed., pp. 444-50.

The points of this space are equivalence classes of points of the standard phase space related by isometries, carrying a geometric structure inherited from that of the standard phase space.[25] Since the standard Hamiltonian is invariant under the group action, it projects down to a well-defined Hamiltonian on the reduced phase space. The resulting dynamical theory captures all of the invariant content of the standard theory.

The reduced phase space is $6N-6$ dimensional, so this space is intermediate in size between the standard phase space and the relative phase space. In fact, we can think of it as being the result of augmenting the relative distances and relative velocities which parameterize the relative phase space with six further variables. These fall into two classes: constants of motion and true dynamical variables. Four of the six are constants of motion: if we fix their initial values, then we are guaranteed that dynamical evolution will leave them fixed. The other two variables are dynamical—like the relative position and velocity variables, we expect dynamical evolution to alter their values.

Three of the four constants of motion are redundant in the following sense: We can stipulate that they take some fixed value and examine the dynamical theory that results. It will consist of a space of states, parameterized by the relational variables, the remaining constant of motion, and the two nonrelational dynamical variables. This space will carry a geometric structure and a Hamiltonian. Neither depend in any way on which value we choose for our three constants of motion. This is just to say that these constants of motion play no dynamical role in our theory—and hence that, from the point of view we are considering, they ought to be dropped from the reduced phase space, yielding a space of dimension $6N–9$. (These three dynamically inert variable describe the velocity of the center of mass of the system, they can be eliminated by identifying points of the original or reduced phase space related by Galileian boosts.)

The remaining nonrelational variables of the reduced phase space are very different. I observed above that merely specifying the relative distances and velocities for our particles is insufficient to determine their future evolution—one also has to know something about the state of rotation of the system that they form. The remaining constant of motion, R, corresponds to the magnitude of the angular momentum of the system, while the two nonrelational dynamical variables serve to fix a point on a sphere of radius R, encoding the direction of the angular momentum of the system in a frame rotating along with it If we fix R, we get a space of variables consisting of the relational variables and our two nonrelational ones The geometric structure and Hamiltonian on this space do depend on the value of R [26]

---

[25] The result is a stratified space, equipped with a Poisson structure adequate for the formulation of classical mechanics. See, for example, Juan-Pablo Ortega and Tudor Ratiu, "Singular Reduction of Poisson Manifolds," *Letters in Mathematical Physics*, XLVI (1998): 359-72.

These three additional nonrelational variables are required to handle the difference between rotating and nonrotating systems of particles—a difference that simply cannot be captured in the strictly relational framework of the relative phase space, which lacks the resources to draw the dynamically crucial distinction between the various possible states of rotation of the system The reduced theory incorporates just enough variables to be able to recognize such differences—without running afoul of PSR by countenancing distinct physical possibilities that are related to one another by symmetries The nine variables included in the standard theory but absent from the reduced theory—the position and velocity of the center of mass, and the orientation of the system about its center of mass—are closely associated with the symmetries of classical mechanics They also play no dynamical role, and it is their elimination which allows one to protect PSR

Let me point out some other features of the reduced theory First, notice that although from the perspective of the standard theory the three remaining nonrelational variables bear a very close relationship to the total angular momentum of the system, we find that from the perspective of the reduced phase space they have no straightforward geometric interpretation Indeed, notice that we never fixed the value of N Thus, we need the same three nonrelational variables in our reduced phase space no matter how large our system is This forces us to conclude that these three variables correspond to properties of the system as a whole, rather than to properties of any particular particle Within the present framework—particles moving in Euclidean space—it is very difficult to see how we can give a geometrically reasonable interpretation of these nonrelational variables [27]

Furthermore, notice that, if $R=0$—if, from the traditional point of view, the angular momentum of the system vanishes—then the two remaining nonrelational dynamical variables drop out of the picture

---

[26] For details, see Toshihiro Iwai, "A Geometric Setting for Classical Molecular Dynamics," *Annales de l'Institut Henri Poincaré*, Section A, 47 (1987) 199-219

[27] Neo-Newtonian space-time provides a more appropriate setting for interpreting the reduced theory, see my "Geometry and Motion "

(since they parameterize a sphere of radius zero) Thus, we end up with a theory on the relative phase space whose variables are just the relative distances and velocities And, in fact, the dynamics of this sector of the reduced phase space coincide with the dynamics for the relative distances and velocities which result if the relational initial data are arbitrarily embedded in the standard phase space, subject to the constraint that the angular momentum vanishes, and their evolution is read off from the evolution of the standard variables Since, as a matter of empirical fact, our universe has no discernible angular momentum, this strictly relational sector of the reduced theory is empirically adequate within the domain of Newton's theory[28]

VII THE GENERAL CASE

The example of the preceding section is typical of one large class of cases We begin with a configuration space for some set of particles or fields, and with a group of symmetries acting on that space, in this case, we expect that a reasonable theory constructed on the phase space of the system will admit that same group of symmetries When this occurs, we can—so long as some relatively weak technical conditions hold—construct an alternative theory on the reduced phase space which arises when we identify points of the original phase space related by symmetries We then find the following (1) The reduced phase space includes variables in addition to those which arise when we begin by quotienting the original configuration space by its group of symmetries, and then augment the resulting set of relative configuration variables with their rates of change These further nonrelational variables of the reduced theory fall into two groups—constants of motion and genuine dynamical variables[29] (2) These nonrelational variables are closely related to the conserved quantities that Noether's theorem associates with the symmetry group of the theory The number of nonrelational variables will be equal to the dimensionality of the group of symmetries In particular, it will be independent of the size of the system under consideration—suggesting that these additional variables ought to be thought of as properties of the system as a whole These variables are often implicated in awkward problems of interpretation (3) Typically, the topology, geometry, and Hamiltonian of the reduced theory will be distinctly more complex than that of the reduced theory Almost always, the reduced Hamiltonian will include potential energy terms corresponding to forces absent in the original theory (4) In the special case corresponding to the vanishing of the conserved quantities associated with our symmetries, we end up with a theory depending only upon the relative configuration and momentum variables When this fragment is empirically adequate, we are able to achieve a fully relational theory

---

[28] This sector very nearly coincides with the theory of Julian Barbour and Bruno Bertotti, "Mach's Principle and the Structure of Dynamical Theories," *Proceedings of the Royal Society of London* A, 382 (1981) 295-306, see my "Geometry and Motion" for references and further discussion

[29] In the special case where the group in question is Abelian, all of the nonrelational variables of the reduced phase space are constants of motion

This gives us some idea of the costs and benefits of saving PSR from symmetries within one large and important family of classical mechanical theories (similar remarks would apply to more general classes of theories) It turns out that devoted advocates of PSR can rest easy—we can (almost) always replace a theory admitting a symmetry group by a mathematically reasonable quotient Others, in light of (3) above, will want to examine carefully the sacrifices required on a case-by-case basis—the reduced theory may appear less than attractive qua physics beside its more familiar symmetric counterpart But one has to be on one's guard here—for judgments about what is natural and what is contrived are strongly conditioned by upbringing and familiarity

It is only relatively recently that the reduction procedure we have been discussing was formulated in its full generality But it has roots in the nineteenth century Heinrich Hertz was very impressed by the possibility—analyzed in detail by Edward Routh—of eliminating those *cyclic variables* whose momenta are irrelevant to a given dynamical problem [30] Hertz hoped that the terms which this process introduces into the reduced Hamiltonian would provide one of the keys to eliminating the notion of force altogether from classical mechanics—he believed that the forces which appear in our theories of macroscopic phenomena are in part an artifact of our lack of knowledge of the true microscopic degrees of freedom [31] We have, he thought, effectively performed a reduction without realizing it—what we think of as being the proper form for a physical theory is in fact highly misleading, the result of an ontologically inappropriate reduction

---

[30] For an elementary discussion, see Cornelius Lanczos, *The Variational Principles of Mechanics* (New York Dover, 1986), §V 4 The standard textbook example is the Kepler problem—the motion of a point particle in a central force field—where the elimination of an angular variable results in a theory with a force which varies inversely with r^3, see V I Arnold, *Mathematical Methods of Classical Mechanics* (New York Springer, 1989), §2 8

[31] *The Principles of Mechanics* (New York Dover, 1956) For a helpful sketch of Hertz's program, see Lanczos, *Variational Principles*, §V 5

(An aside about Hertz's program The idea was to start with a flat configuration space, then to reconstitute physics as we know it through the imposition of constraints on phase space and the elimination of cyclic variables All of physics would then follow from (Euclidean) kinetic energy and rigid connections between particles—force would play no role The program is sometimes dismissed by philosophers as hopeless—What mechanism of hidden microscopic degrees of freedom could account for the gravitational force? But, in fact, it is all too easy to implement technically[32] The problem is rather that known implementations have little discernible *physical* interest The kinetic energy in question bears no straightforward relation to any physical notion[33] Even worse for almost any physical system, it suffices to add a *single* hidden cyclic degree of freedom)

In a similar vein, our theory of the behavior of a charged particle in an electromagnetic field—which has the particle tracing out a curved trajectory in four-dimensional space-time in accord with the Lorentz force law—strikes us as being entirely natural But this theory can be thought of as arising via reduction from a Kaluza-Klein theory, in which space-time has a fifth (minuscule and circular) dimension and the charged particle moves along a straight line[34] Here the question of which is more natural or attractive is a difficult one—and we might well expect our intuitions to be hostage to the success of Kaluza-Klein-inspired strategies in string theory and elsewhere

My point is this while it is quite true that we have no guide here other than the sum-total of our judgments of simplicity, elegance, and so on, we must be willing to interrogate this guide rather carefully, and with as much self-consciousness as we can muster

VIII METHODOLOGICAL WORRIES

Very often, then, when presented with a physical theory invariant under a group of symmetries, we have the option of replacing it by a reduced theory with the same power Ought we to do so? It is often asserted that symmetries are crucial in modern physics they underlie conservation laws, they enforce relativity principles, they guide theory construction Can we afford to sacrifice these virtues?

---

[32] Almost any finite-dimensional classical theory can be handled using some sequence of standard tricks and results Handy ingredients include Theorem 3 15 of Arnold, V Kozlov, and A Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics* (New York Springer, 1997), Theorem 2 of Mark Gotay and Gijs Tuynman, "R^{2n} is a Universal Symplectic Manifold for Reduction," *Letters in Mathematical Physics*, xviii (1989) 55-59, the trick deployed on p 43 of Guillemin and Sternberg, *Variations on a Theme by Kepler* (Providence, RI American Mathematical Society, 1990), the embedding theorems of Mikhael Gromov, *Partial Differential Relations* (New York Springer, 1986), §§3 1 and 3 3

[33] I am here indebted to Howard Stein's criticism of a related program of Henri Poincaré in his unpublished paper "Physics and Philosophy Meet The Strange Case of Poincaré "

[34] In this case, charge appears as the momentum of the particle along the fifth dimension, see, for example, Marsden, *Lectures on Mechanics*, §§7 5 and 7 6

There is much to be said here But the short answer is that no real sacrifices are required Conservation laws are enforced in the reduced theory, just as in the original theory—indeed, they are encoded in the geometry of the reduced phase space in a rather interesting way As for relativity principles and other heuristics, note that there is a sense in which the symmetry group is still present, in a more subtle form by looking at the local geometry of the space of states, one can divine which group, if any, was factored out in its construction [35] So nothing need be lost in moving to the reduced theory

But there is something to be gained Invariance of a theory under (continuous groups of) symmetries implies that the phase space of the theory includes dynamically inert variables (since the action of the group on a point of phase space carries us to a point where the dynamics look exactly the same) By moving to the reduced phase space, we avoid the sort of redundancy in our space of possibilities which underwrote Clarke's attack on PSR, without losing any of the (invariant) content of our dynamics

IX PSR IMPLIES PII?

I want to discuss briefly one final topic the logical relationship, within the present context, between my version of PSR and Hacking's version of PII I mentioned above that the construction of a reduced phase is contingent upon some technical conditions One that is often imposed requires that the identity be the only symmetry that has any fixed points This guarantees that the reduced phase space will be a manifold More generally, this will follow so long as all of the points of our original phase space manifest the same degree of symmetry [36] But when this fails, the reduced phase space will not be a manifold generic points of the original space, corresponding to minimally symmetric arrangements of matter, will project down to generic, well-behaved points of the reduced phase space, but the reduced phase space will also have singularities that correspond to those points of the original space which represent more symmetrical arrangements of matter

---

[35] This can be made precise, see Alan Weinstein, "The Local Structure of Poisson Manifolds," *Journal of Differential Geometry*, XVIII (1983) 523-57

[36] That is, we require that the isotropy subgroups associated with each point of our phase space be conjugate to one another within the symmetry group

Now, this seems to mean that enforcing PSR (in my sense) implies PII (in Hacking's sense), since if one wants a nonsingular reduced phase space, one must exclude symmetric arrangements of matter from one's original space of physical possibilities. And this seems meet and right, since Leibniz often proclaimed that his entire philosophy, including lesser principles such as PII, rested ultimately upon the Principles of Noncontradiction and Sufficient Reason.[37]

Alas, the technical situation is more complex than this. It is not so much that the reduced theory makes no sense near singularities—the reduced phase space is well-behaved topological space, carrying enough geometric structure to underwrite dynamics—as that one must be more careful in formulating it, since the structure of the tangent space of the reduced phase space changes discontinuously as one moves from generic points to singular points.[38] But perhaps one could still cobble together a Leibnizean argument from PSR to PII via the Law of Continuity (which might be thought to prohibit such jumps)?[39] Or perhaps this last anachronistic epicycle is one too many.

GORDON BELOT

New York University

---

[37] See, for example, L II 1 and L V 21. On this issue, see Cover and O'Leary-Hawthorne, chapter 5.

[38] This picture has been established for finite dimensional systems (in addition to Ortega and Ratiu, "Singular Reduction," see Reyer Sjamaar and Eugene Lerman, "Stratified Symplectic Spaces and Reduction," *Annals of Mathematics*, CXXXIV (1991): 375-422), and for some important infinite dimensional systems (James Isenberg and Marsden, "A Slice Theorem for the Space of Solutions of Einstein's Equations," *Physics Reports*, LXXXIX (1982): 179-222; Jędrzej Śniatycki, Günter Schwartz, and Larry Bates, "Yang-Mills and Dirac Fields in a Bag, Constraints and Reduction," *Communications in Mathematical Physics*, CLXXVI (1996): 95-115).

[39] Leibniz deployed this Law in rejecting laws of nature proposed by Descartes and Malebranche; see Loemker, ed., pp. 351-53 and 397-403.

CAUSES AS PROBABILITY RAISERS OF PROCESSES*

Causation, according to David Hume, is one of the three fundamental conceptual relations (along with resemblance and contiguity), and is the foundation of all reasoning concerning matters of fact. Causation, according to various contemporary philosophers, is required for the analysis of metaphysical concepts such as persistence, scientific concepts such as explanation and disposition, epistemic concepts such as perception and warrant, ethical concepts such as action and responsibility, legal concepts such as homicide and negligence, mental concepts such as functional role and conceptual content, and linguistic concepts such as reference, to name just a salient few. Yet the nature of the causal relation itself remains mysterious.

The leading accounts of the nature of causation divide into *probability-raising* and *process-linkage* views. On the probability-raising view, causation is rooted in the comparative probability of the effect with the cause versus without. On the process-linkage view, causation is rooted in the existence of a connecting line from cause to effect. I propose a third alternative which synthesizes these views while solving their problem cases. On this alternative, causation is rooted in the comparative probability of the connecting line to the effect with the cause versus without: causes as probability raisers of processes.

In the first two sections, I (briefly) outline the probability-raising and process-linkage views and identify a space of problems, and then I focus on developing the alternative probability-raisers-of-processes solution. I identify problems for this view as well, though I conclude that it is at least in important respects a step forward.

I. PROBABILITY-RAISING AND PROCESS-LINKAGE VIEWS

The probability-raising view of causation has its roots in the social sciences and in the philosophical analysis of causation in terms of necessary and/or sufficient *conditions*. In the social sciences, issues concerning causation are addressed by the use of algorithms that seek covariations in frequency data, and then sort the causal covariations

---

* Thanks to David Coady, Cian Dorr, Harty Field, Ned Hall, Chris Hitchcock, Barry Loewer, Tim Maudlin, Michael McDermott, Brian McLaughlin, Ram Neta, L. A. Paul, and especially to David Lewis.

0022-362X/01/9802/75–92

from the merely correlational ones.[1] Many social scientists have operationally identified the causal relation with these algorithms. This is a statistical conception of causation.

The probability-raising view also represents the indeterminism-induced synthesis of philosophical accounts of causes in terms of necessary and/or sufficient conditions for their effects. The natural extension of the necessary condition view into the indeterministic case is that causes are necessary, not for the presence of their effects (since the effect might still retain some nonzero chance of spontaneous eruption even without the cause), but for a higher probability of their effects. Thus, for David Lewis,[2] causation is rooted in the counterfactual claim that, had the cause not occurred, then the chance of the effect would have been less, and D. H. Mellor[3] requires that $\mathrm{Ch}_C(E) > \mathrm{Ch}_{\neg C}(E)$, which measures the chance that the actual circumstances give E versus the counterfactual chance that the circumstances minimally changed with respect to C would give E. The natural extension of the sufficient condition view into the indeterministic case is that causes are sufficient, not for the presence of their effects (since the effect might still retain some nonzero chance of fizzling even with the cause), but for a higher probability of their effects. Thus, for Patrick Suppes,[4] causation is rooted in the statistical claim that the limiting frequency of effect-type events is greater with cause-type events than without: $\mathrm{p}(E|C) > \mathrm{p}(E|\neg C)$, and Ellery Eells[5] requires that the chance of E increase to a high value just after the time of C and remain high until the time of E. And so counterfactual dependence flows into counterfactual chance dependence, constant conjunction flows into comparatively frequent conjunction, and the two merge into the wider seas of probability raising.

The process-linkage view of causation has its roots in physics and in the philosophical analysis of causation in terms of physical connections rather than abstract conditions. Physicists have long suggested that flows of energy momentum, or something of that sort, are the basis for actual causal connections.[6] Many philosophers have rejected the necessary and/or sufficient conditions approach to causation in favor of physical connections David Fair[7] identifies actual causation with energy flow, Wesley Salmon[8] describes causal processes as those connections capable of propagating their own structure (transmitting a mark), and Douglas Ehring[9] requires the persistence of a trope from cause to effect Think of the path of energy flow, the wave of structural propagation, and the worldline of the persisting trope as constituting a physical connecting line from cause to effect

---

[1] Peter Spirtes, Clark Glymour, and Richard Scheines offer a sophisticated algorithmic approach in their *Causation, Prediction, and Search* (New York: Springer, 1993).

[2] "Causation," in *Philosophical Papers*, Volume II (New York: Oxford, 1986), pp. 172-213.

[3] *The Facts of Causation* (New York: Routledge, 1995).

[4] *A Probabilistic Theory of Causality* (Amsterdam: North Holland, 1970).

[5] *Probabilistic Causality* (New York: Cambridge, 1991).

[6] Wladyslaw Krajewski—"Energetic, Informational, and Triggering Causes," *Erkenntnis*, XLVII (1997): 193-202—traces energy flow accounts of causation back to Robert Mayer (the discoverer of the law of conservation of energy) in 1842.

I propose to separate the questions of how the probability-raising and process-linkage relations are best analyzed (which I have discussed at length elsewhere), from the question (which I focus on here) of how the causal relation should then be analyzed In the remainder of this section, I sketch my preferred analyses of probability raising and process linkage, and in what follows I focus on the analysis of causation

Probability raising is best analyzed, I think, in terms of Lewis's counterfactual chances, on which *C* is a probability raiser of *E* if and only if (i) *C* and *E* are actual distinct events, (ii) $\mathrm{ch}(E)\text{-at-}t_C=p$, and (iii) $\neg C \,\Box\!\!\rightarrow \mathrm{ch}(E)\text{-at-}t_{\neg C}<p$ In saying that *C* and *E* are actual distinct events, I mean that each is a ⟨Property, Region⟩ ordered pair whose first member is actually instantiated at its second, with the *C* and *E* properties logically independent and the *C* and *E* regions spatiotemporally disjoint In saying that $\mathrm{ch}(E)\text{-at-}t_C=p$, I mean that the propensity (objective single-case chance) of *E* in the circumstances at the time of *C* is *p* In saying that $\neg C \,\Box\!\!\rightarrow \mathrm{ch}(E)\text{-at-}t_{\neg C}<p$, I mean that the counterfactual supposition of the nonoccurrence of *C* entails that the propensity of *E* in those same circumstances with *C* deleted would have been less than *p*[10]

Process linkage is best analyzed, I submit, in terms of pairwise nomic subsumptions processes are lawful sequences[11] Not just any lawful sequence will do. The antecedent must be prior to the consequent, and the lawful connection must not be revealed to be merely correlational by a more fundamental law. More precisely, C is *directly* process linked to E if and only if (i) C and E are actual distinct events, (ii) there is a law L of the form $(\forall R)\ (\forall R')\ (\langle P, R\rangle \Box \langle P', R'\rangle)$ such that C is an instance of the antecedent and E of the consequent, (iii) there is no law L' more fundamental than L such that actual prior event A is an instance of the antecedent and $C\&E$ is an instance of the consequent, and (iv) $t_C < t_E$. In speaking of laws as having the form $(\forall R)\ (\forall R')\ (\langle P, R\rangle \Box \langle P', R'\rangle)$, I take the quantifiers to range over regions, and the laws to relate the properties of one region to those of another. Thus events, understood as $\langle P, R\rangle$ pairs, can instantiate the antecedents and consequents of laws. Process linkage in general is then defined from the direct case: C is process linked to E if and only if there is a chain of direct process links between C and E.[12]

---

[7] "Causation and the Flow of Energy," *Erkenntnis*, XLVIII (1979) 219-50

[8] *Scientific Explanation and the Causal Structure of the World* (Princeton University Press, 1984) In his *Causality and Explanation* (New York Oxford, 1998), Salmon responds to a suggestion from Phil Dowe (for whom causal processes are the worldlines of conserved-quantity-bearing objects—"Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science*, LIX (1992) 195-216), by moving to a conserved-quantity transmission view of causal processes This remains squarely within the process-linkage camp

[9] *Causation and Persistence* (New York Oxford, 1997)

[10] For further discussion, see Lewis as well as Mellor

[11] David Armstrong proposes "A law (an entity in the world) provides the ground in nature for an inference The cause is like the antecedent"—"The Open Door Counterfactual versus Singularist Theories of Causation," in H Sankey, ed , *Causation and Laws of Nature* (Boston Kluwer, 1999), pp 175-85, here pp 183-84

So when a brick is thrown through a window, on the lawful sequence analysis of *process*, there is a sequence of events from the throwing of the brick, through its intermediary trajectories, to the shattering of the window, and the events in this sequence will be pairwise lawfully related (and not merely covariants, and rightly temporized). The initial conditions of the brick being *here* with a certain momentum at t_0 will lawfully entail that the brick flies to *there* at t_1. And since energy flows are a this-worldly instance of lawful evolution, the lawful subsumption analysis both undergirds the idea of, and inherits the plausibility of, Fair's empirical identification of processes with energy flows.

I claim that however the probability-raising and process-linkage relations are interpreted, causation cannot be understood by probability raising or process linkage alone, but rather should be understood by a synthesis of these relations on which causes are probability raisers of processes. Since the discussion that follows is relatively independent of any specific understanding of probability raising and process linkage,[13] for the sake of generality, I speak of 'probability raising' and 'process linkage' without further interpretation (unless otherwise indicated). The reader is invited, though not required, to understand these notions as I do.

---

[12] For more details, see my "Processes as Lawful Sequences" (unpublished).

[13] I do assume that (i) both the probability-raising and process-linkage relations are asymmetric, (ii) the probability-raising relation is circumstance sensitive, and (iii) the process-linkage relation extends to worlds with different laws than ours (and so is not tied into any contingent details of actual physics).

Supposing we have some notion of probability raising and some notion of process linkage in hand, how (if at all) can causation be analyzed?

II. THE PROBLEM SPACE

Causation cannot be analyzed purely in terms of probability raising or purely in terms of process linkage, however these notions are interpreted and however they are refined. There are counterexamples to these approaches which extend to even their most sophisticated refinements. These counterexamples interlock to form a problem space that no extant account of causation can navigate.

(1) Causation without probability raising (*preemption*, and its *trumping* variant). Pam throws a brick through the window. Meanwhile, Bob (a more reliable vandal), holds his throw on seeing Pam in action, though had Pam not thrown Bob would have.[14]

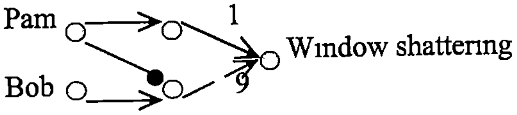


DIAGRAM 1

Preempting Pam's throw is obviously a cause of the window shattering. But her throw is a probability lowerer of the shattering: since Bob is a more reliable vandal, the window's chances would have been worse with Bob in action. Thus probability raising is not necessary for causation.[15]

Probability-raising theorists are well aware of preemption and have responded with an extensive literature of refinements. These refinements divide into two general strategies: *factor intermediaries* and/or *require precision*. Intermediary-based approaches look to intermediary events or probability evolutions to try to find a chain of causations from preemptor to effect, and precision-based approaches look to the exact time and manner of the events involved to try to find differences in the effect attributable to the preemptor.[16]

---

[14] Diagram conventions: filled circles doubly represent neurons that fire and events that occur, unfilled circles doubly represent neurons that do not fire and events that do not occur, arrows represent stimulatory connections, dots inhibitory connections, and numbers probabilistic connections of varying strengths. If two neurons are connected by a nonprobabilistic stimulatory connection and the first fires, the second will fire unless inhibited.

[15] For further discussion of these cases, see Lewis; I. J. Good, "A Causal Calculus I," *British Journal for the Philosophy of Science*, XI (1961): 305-18; and Peter Menzies, "Probabilistic Causation and Causal Processes: A Critique of Lewis," *Philosophy of Science*, LVI (1989): 642-63.

There are, however, variants of preemption (*trumping*) with no intermediaries to factor or differences to be precise about, which immediately undercut all extant probability-raising-based strategies against preemption.

The silver-tongued sergeant and mumbling major both order 'Charge!' and so the corporal decides to charge. The major's order is a cause of the corporal's decision, as can be seen from the fact that ranking orders trump: when orders conflict the corporal will follow the major, so when they accidentally converge the same should hold. But the major's order is a probability lowerer of the corporal's decision: the probability of the corporal deciding to charge in those circumstances would have been greater had the major left it to the silver-tongued sergeant to give the order since the major is more likely to be misheard.[17]

It is an empirically open question (concerning exactly how the mind works) whether there are any relevant intermediaries between order and decision, and/or any precisifiable differences in the decision attributable to the major: certainly, the corporal need not be conscious of either. And so it is possible to have preemption without intermediaries to factor or differences to be precise about.[18]

---

[16] Examples of intermediary-based approaches include Lewis's original approach which takes the transitive closure of probability-raising, Menzies's requirement of temporally continuous chains of probability raisings, Eells's use of probability trajectories through the (C, E) temporal interval, and Igal Kvart's constructions based on further conditionalization on temporally intermediate events. These intermediary-based approaches are offered, respectively, in Lewis, Menzies, Eells, and Kvart, "Cause and Some Positive Causal Impact," in J. Tomberlin, ed., *Philosophical Perspectives 11: Mind, Causation, and World* (Cambridge: Blackwell, 1997), pp. 401-32.

Examples of precision-based approaches include Lewis's fragility proposal on which probability raising is assessed with C and E required to have their exact actual times and manners, L. A. Paul's time-tracking emendment on which C's being a probability raiser of E is disjoined with C's being a probability raiser of E-at-t; and Deborah Rosen's suggestion that probability raisings that disappear under precisification are spurious. These precision-based approaches are offered, respectively, in Lewis, Paul, "Keeping Track of the Time: Emending the Counterfactual Analysis of Causation," *Analysis*, LVIII (1999): 191-98, and Rosen, "Discussion: In Defense of a Probabilistic Theory of Causality," *Philosophy of Science*, XLV (1978): 604-13.

[17] The probabilistic impact buttresses the causal claim, since the chance of the corporal deciding to charge is keyed into the chance of him correctly understanding the major's mumbling (and is independent of the chance of him grasping the sergeant's order).

(2) Probability raising without causation (*fizzling*, and its *overlapping* variant). Fred grabs a brick and aims. Pam, oblivious to Fred, throws a brick through the window. Perhaps she preempts Fred, or perhaps Fred quits on his own, or perhaps he throws but wide. Suppose that at the time Fred aimed there was some non-one chance of Pam succeeding and some nonzero chance of Fred succeeding without Pam. (The preemption diagram above represents just such a case, with Bob as the fizzler.)

Fizzled Fred's aiming is obviously not a cause of the window shattering. But his aiming is a probability raiser of the shattering, given Pam's non-one and Fred's nonzero chance of success at the time: when Fred aimed the window's prospects were worse. Thus probability raising is not sufficient for causation.[19]

Probability-raising theorists have only recently become aware of fizzlings, but the same general strategies of factoring intermediaries and/or requiring precision suggest themselves. There are, however (as with preemption), variants of fizzling (*overlapping*) with no intermediaries to factor or differences to be precise about.

Suppose that an atom of U238 and of Ra226 are placed in a box at t_0, and that at t_1 there is an atom of Th234, an alpha particle, and (still) an atom of Ra226 in the box. The presence of the radium at t_0 is not a cause of the presence of an alpha particle at t_1. Since Ra226 decays into Rn222 plus an alpha particle, it is evident from the absence of Rn222 (and the continuing presence of the Ra226) that the alpha particle was not produced by radium decay, but rather by the independent process of uranium decay in which U238 decays into Th234 plus an alpha particle. But the presence of the radium is a probability raiser for there being an alpha particle, as follows from the lawful probability λ' of radium decay.

Since particle emissions from radioactive sources (as standardly understood) occur spontaneously and directly, there will be no hidden intermediaries to factor, and if at least one of the particles is in a superposition of location so that the two are "smeared" over a common region, then there can be no differences at all in the alpha

---

[18] For a full discussion of these cases, see my "Trumping Preemption," this JOURNAL, XCVII, 4 (April 2000): 165-81.

[19] For more on these cases, see Menzies; and Christopher Read Hitchcock, "Do All and Only Causes Raise the Probabilities of Effects?" (unpublished).

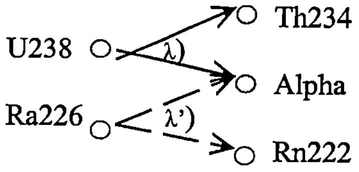


DIAGRAM 2

emission projected from either source So no manner of factoring intermediaries and/or requiring precision will help Thus probability raising, even when refined to factor intermediaries and/or require precision, is not sufficient for causation[20]

So far, it seems as if the causal relation tracks the process-linkage relation rather than the probability-raising relation The preempting cause is connected to the effect and the fizzled noncause is not (just look at the diagrams!) But there are also cases where the causal relation looks to track the probability-raising relation rather than the process-linkage relation

(3) Causation without process linkage (*disconnections*) the plane is on a collision course with the mountain The air traffic controller is about to alert the pilot to turn, when the saboteur destroys the radio tower, and so the alert is not transmitted, and so the plane crashes

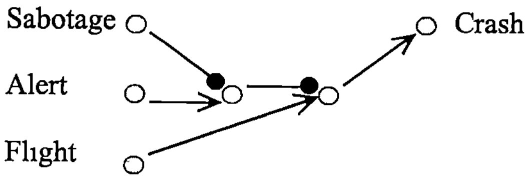


DIAGRAM 3

The sabotaging is obviously a cause of the crash, which intuition may be buttressed by the following considerations Statistically, sabotages in these circumstances will be universally followed by crashes Counterfactually, had the tower not been sabotaged then the crash would not have occurred The explanation for why the plane crashed will surely include the sabotage Knowing that the sabotage occurred will license an inference to the crash Sabotages in these circumstances are an effective means for crash-hungry agents The saboteur will be held morally responsible for the crash In short, all the connotations of causation are in full force

---

[20] For further discussion, see my "Overlappings Probability-Raising without Causation," *Australasian Journal of Philosophy*, LXXVIII (2000) 40-46

But as the diagram suggests, no process links sabotage to crash No relevant energy-line/structure/trope or whatnot persists between them, and no lawful sequence links them Rather, what is salient here is the *absence* of a process If only there had been an alert transmitting process connecting tower to pilot, the crash might well have been prevented, but the sabotage prevented that [21]

Process-linkage theorists appear to be largely unaware of these cases, or at least unaware of how ubiquitous and paradigmatically causal they are Some process-linkage theorists have (for independent reasons) refined their semantics by *covering absences* and/or *conjoining probability raising* absence covering approaches add counterfactuals about processes to cover the case of cause and/or effect being an absence ($\neg C/\neg E$), and probability-raising conjunctions require process linkage and probability raising [22] But covering absences will not help since the disconnector and effect are both presences, and conjoining probability raising (or anything else) will not help since we already have causation without process linkage Thus process linkage, even when refined to cover absences and/or to require probability raising, is not necessary for causation

(4) Process linkage without causation (*traces*) Pam throws the brick through the window while innocent Tom watches in dismay, or eats a sandwich, or contemplates fate Obviously, there is no causation

---

[21] Disconnections are present in even the most paradigmatic causal sequences Gus gets mad (C), pulls the trigger (D_1), fires a bullet through Vic's heart (D_2), and Vic dies (E) Here is causation if anything is But if one examines how muscles contract, how guns fire, and how heart injuries kill, one will find disconnections at every step For instance, heart injuries kill not by a connecting process, but rather by disconnecting what would otherwise have been a process, namely, the pumping of oxygenated blood to the brain For further discussion, see Ned Hall, "Two Concepts of Causation" (unpublished), Lewis, "Void and Object" (preprinted by the Department of Philosophy, University of Melbourne, 1998), and my "Causation by Disconnection" (forthcoming in *Philosophy of Science*)

[22] Fair and Dowe both offer absence-covering extensions (although Dowe relegates the extension to the ersatz status of "causation*"), and Ehring as well as Hector-Neri Castañeda both conjoin the process-linkage and probability-raising approaches These refinements appear, respectively, in Fair, Dowe, *Physical Causation* (unpublished), Ehring, and Castañeda, "Causes, Causity, and Energy," in Peter French, Thomas Uehling, Jr , and Howard Wettstein, eds , *Midwest Studies in Philosophy, Volume IX Causation and Causal Theories* (Minneapolis Minnesota UP, 1984), pp 17-28

from Tom's watching or whatever else to the window shattering if ever there is a lack of causation, it is here. But there are presumably lots of processes connecting Tom to the shattering. Photons and sound waves and dust motes and other *mere traces* of Tom's watching connect these events.

The obvious thing to say about traces is that they are *irrelevant* connections. But relevance constraints (which counterfactuals express so naturally) are just what the process-linkage semantics foregoes.[23]

Covering absences clearly will not help since *C* and *E* are presences here. Conjoining probability raising will help (assuming the probability-raising relation in use does not require complete precision), but at the cost of newfound failure in preemption cases since, as shown above, preemptions involve causation without probability raising. And conjoining probability raising will not even secure sufficiency if the trace comes from a fizzler: fizzled Fred's aiming is *both* a probability raiser of the window shattering *and* (presumably) connected to the shattering by various stray photons, but is still not a cause. Thus process linkage, even when refined to cover absences and/or to require probability raising, is not sufficient for causation.

In disconnections and traces it seems as if the causal relation tracks the probability-raising relation rather than the process-linkage relation. The disconnector is a probability raiser while the trace noncause is not (the sabotage makes the crash more likely while the stray photon does not make the window shattering more likely).

The four problem cases rendered above interlock to form a single problem space that no extant account of causation can navigate:

> *Problem Space*: the preempting cause (1) and trace noncause (4) are both process linked to, but not probability raising of, the effect in question. The disconnecting cause (3) and fizzled noncause (2) are both probability raising of, but not process linked to, the effect in question.

An adequate account of causation must explain the causally relevant difference between preemptors and traces, and between disconnectors and fizzlers (vandals and saboteurs are causes, while failed vandals and bystanders are not) Neither the probability-raising nor the process-linkage relation, however interpreted and however refined, can reveal the nature of causation on their own

---

[23] Salmon, whose theory of explanation was subjected to a related critique by Hitchcock (in "Discussion: Salmon on Explanatory Relevance," *Philosophy of Science*, LXII (1995): 304-20), has conceded this point for explanation: "As a result of Hitchcock's analysis, I would now say that connecting causal processes, in the absence of statistical relevance relations, also lack explanatory import"—"Causality and Explanation: A Reply to Two Critiques," *Philosophy of Science*, LVIX (1997): 461-77, here p. 476. I would add that, given how entwined causation and explanation are (especially for Salmon), the same should hold for causation.

III PROBABILITY RAISERS OF PROCESSES

Causation, I here propose, is rooted in the comparative probability of the connecting line to the effect with the cause versus without This approach synthesizes the probability-raising and process-linkage views and successfully navigates the problem space mapped above.

For starters, consider

> *Analysis 1* C causes E if and only if C is a (P)robability (R)aiser (O)f a (P)rocess for E (for short C is a PROP for E)

The idea is that, to assess if C causes E, first identify the actual process(es) linked to E, and second assess the probabilistic relevance of C, not to E, but to the comparatively extended event that comprises the E-process (or one of them in *overdeterminations* with multiple processes) The E-process, for a given candidate cause C, is to be traced back to the time of C, and may or may not include C In worlds in which the causal order is the temporal order, the E-process runs the closed interval $[t_C,t_E]$, and in most actual world cases the E-process may be thought of as the energy lines in $[t_C,t_E]$ converging on E (an E-process is thus *not* the "causal intermediaries" between C and E)

Given the counterfactual chance understanding of probability raising and the nomic subsumption understanding of process linkage sketched above, Analysis 1 may be further interpreted as

> *Analysis 1 Interpreted* C is a PROP for E if and only if (i) there is an extended event 'E-line' containing actual distinct events $\langle C', D_1, D_2,\ \ , D_n, E\rangle$ in pairwise nomic subsumption relations, (ii) there is an actual event C at t_C' which is distinct from D_1, D_2, , D_n, and E (C may or may not be distinct from C'), (iii) ch(E-line)-at-$t_C=p$, and (iv) $\neg C \ \Box\!\!\rightarrow$ ch($\langle E$-line$\rangle$)-at-$t_{\neg C}<p$

This interpretation of Analysis 1 is compatible with a purely "Humean" metaphysic of just occurrent properties and space-time, given a prior Humean account of laws, chances, and counterfactuals Of course, other interpretations are available even the causal primitivist may accept Analysis 1 as revealing how the notions of *probability* and *process* derive from primitive causation

Analysis 1 synthesizes the probability-raising and process-linkage views Analysis 1 may be thought of as a development of the precision strategy for the probability-raising view Lewis considers precision with respect to time, precision with respect to manner, and precision with respect to causal origins. He rejects precision with respect to time and/or manner as oversensitive to spurious influences, and rejects precision with respect to causal origins as trivializing the account of causation. But there is a fourth possibility (which is neither oversensitive nor trivializing): precision with respect to lawful origins. If we think of *E* as having its lawful origins essentially, then *E* exists if and only if *E*-line exists. It would then follow that *C* is a probability raiser of *E*-line if and only if *C* is a probability raiser of *E*-with-its-actual-lawful-origins.[24]

Analysis 1 may also be thought of as an extension of the absence-covering strategy for the process-linkage view. Fair, who saw that absences can be causal, extended his view by considering counterfactuals about processes:

(1) *C* causes *E* if and only if (a) *C* and *E* occur, and (b) *C* is physically connected to *E*.
(2) *C* causes $\neg E$ if and only if (a) *C* occurs and *E* does not, and (b) *C* is physically connected to an E' incompatible with *E*.
(3) $\neg C$ causes *E* if and only if (a) *C* does not occur and *E* does, and (b) had *C* occurred it would have been physically connected to an E' incompatible with *E*.
(4) $\neg C$ causes $\neg E$ if and only if (a) neither *C* nor *E* occurs, and (b) had *C* occurred it would have been physically connected to *E*.

In the case where *E* is an absence, think of the E' incompatible with *E* as the *manifestation* of *E*'s absence: if *E* is the absence of beer in the fridge, its manifestation is the fridge actually being stuffed with milk, grapes, and air. If we think of the ⟨Property, Region⟩ pair of absent *E*s as ⟨Manifestation, Region⟩, then Analysis 1 can be seen as both a unified generalization over, and indeterministic extension of, Fair's four cases, all of which can now be seen to relate how the *E*-manifestation process depends on whether or not *C* occurs.

Thus Analysis 1 is a true synthesis of the probability-raising and process-linkage approaches, in that it not only explicitly incorporates both notions, but also extends the leading refinements of these approaches to convergence.

Analysis 1 also explains the success of counterfactual accounts of causation. In the deterministic case, if *C* and *E* both occur and *E* counterfactually depends on *C*, then it follows that whatever process actually produced *E* must depend on *C* (else *E* would still occur absent *C* via this independent process), so *C* is a PROP for *E* In the other direction, if *C* is a PROP for *E* and there is no other *E*-process (no redundant causation) then *E* will counterfactually depend on *C* Thus Analysis 1 explains why, at least in the deterministic case, the counterfactual account of causation proves such an excellent account, except for the seemingly intractable difficulties created by redundant causation (overdetermination and preemption)

---

[24] Analysis 1 is still preferable to lawful-origin precision, since the latter makes dubious claims about event essences (though perhaps the essentiality of lawful origins of events is defensible on analogy with the essentiality of parental origins of persons), and obscures just how much of the work of the theory consists in spelling out an appropriate account of lawful origins.

Analysis 1, moreover, yields a unified and informative solution to the problem space rendered above Consider a given candidate cause *C* If the *E*-process includes *C*, then *C* is a PROP for *E* if and only if *C* is an essential part of the *E*-process without *C*, if *E* still occurs it is via a different process entirely, rather than the same process slightly altered If the *E*-process does not include *C*, then *C* is a PROP for *E* if and only if *C* is a shield for the process without *C*, the chance of that same process running to completion would have been less Analysis 1 is satisfied by essential parts of, as well as shields for, processes With this the problem cases above all fall into line

In preemptions, the preemptor comes out a cause because it is an essential part of the *E*-process without the preemptor, even though *E* still occurs it is by a different process entirely, namely, by the backup process In the case of Pam and Bob, the window-shattering process (assessed from the time of Pam's throw) is the sequence of events of her throwing the brick, her brick flying ever closer to the window, hitting it, and producing waves of force that shatter it Pam's throw is a part of this process and an essential part, since without it if the window gets shattered at all it will be by a different process entirely Bob's process This solution extends to trumpings, given a suitably general conception of process to appreciate that the trumper (major) is on the process to the effect and the backup (sergeant) not [25]

Traces, however, come out noncauses because they are inessential parts of the *E*-process without the trace, *E* still occurs by the same process slightly altered Thus, without the traces of Tom's watching, the process by which Pam throws her brick through the window would be in essence the same, minus a few irrelevant stray photons [26]

---

[25] Applying the lawful sequence analysis of processes there are trumping laws linking ranking orders to decisions Since the major's order is the ranking order, only the major's order, not the sergeant's, instantiates the antecedent The corporal's decision instantiates the consequent

[26] We get along perfectly well without a defined counterpart relation for events, persons, tables, and the like, so in this respect processes are on par with the other furniture of the world Our robust intuitions serve us sufficiently well

There is a continuum of cases between traces and genuine causes (just imagine the dust mote from Tom swelling until it is as impactful as Pam's brick), and obviously we cannot determine at which point noncause turns into cause Likewise, there is a continuum of cases between the same process slightly altered and a different process entirely So our understanding of the causally relevant difference between preempting causes and trace noncauses is vague along exactly the same continuum as our understanding of the counterpart relation for processes analysans swings with analysandum Thus Analysis 1 not only explains the causally relevant difference between preemptions and traces (essential versus inessential process parts), but does so in the right way to mirror the logical structure involved

In disconnections, the disconnector comes out a cause because it shields the *E*-process from what otherwise might have interrupted it without the disconnector, the chance of the *E*-process running to completion would have been less In the sabotage case, the crashing process consists in the main of the plane flying mountainward through the relevant interval and then crashing Without the sabotage, the chance of this process would have been less, because the alert would then have been transmitted and the pilot would then have turned the plane away Thus the sabotage shields the crashing process from the alert

Fizzlers, in contrast, come out noncauses because they are neither essential parts of, nor shields for, the *E*-process Thus Fred's misfiring is neither a part of nor a shield for the process by which Pam throws her brick through the window This solution extends to overlappings, given a suitably general conception of process to appreciate that the cause (U238) is connected to the effect and the noncause (Ra226) not[27]

IV REFINEMENTS

Analysis 1, which identifies causes with probability raisers of processes, while an improvement over its predecessors with respect to the space of problems discussed above, is subject to (at least) three objections First, the account fails to provide for the *transitivity* of causation There are cases (example to be given shortly) in which *C* is a PROP for *D* and *D* is a PROP for *E*, but *C* is not a PROP for *E* Second, the account fails to provide for the *continuity* of causation It seems that, at least in the actual world (leaving the mysterious EPR correlations of quantum mechanics out of the discussion), all causation involves spatiotemporally continuous chains of causal intermediaries But there are actual-world-ish cases (example to be given shortly) in which *C* is a PROP for *E* but there are gaps in the line of PROPs from *C* to *E* Third, the account is subject to counterexample from preemptive disconnections

---

[27] Applying the lawful sequence analysis since the decay laws are U238 → (Th234 & alpha) and Ra226 → (Rn222 & alpha), the absence of Rn222 shows that the consequent of the radium decay law is not instantiated Thus the process in question is not radium decay Rather, as the presence of Th234 shows, the process in question is uranium decay

All three objections can be brought out in the following case, which feeds a preemption case into a disconnection (*preemptive disconnection*) Pam and Bob are out to sabotage the radio tower and prevent an alert from reaching the pilot Pam places her puny bomb (1 chance of blowing up the tower), while Bob is too overcome with laughter to proceed (he would have almost certainly blown up the tower had Pam not distracted him) Pam's bomb then actually succeeds in destroying the tower (preemption), and so the alert is prevented, and so the plane crashes (disconnection)

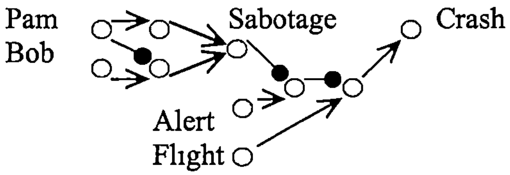


DIAGRAM 4

Pam's saboteuring is clearly a cause of the crash and Bob's is not, but Analysis 1 unfortunately gets this backward Pam's saboteuring is a probability lowerer of the crash process (which consists, in the main, of the plane flying mountainward through the relevant interval) since Bob is a more reliable saboteur, the plane's chances would have been worse with Bob in action And Bob's saboteuring is a probability raiser of the crash process, given that there was some chance he might have managed not to get distracted by Pam In short, Pam is a preempting shield and Bob a fizzled shield for the crash

It is worth noting that preemptive disconnections are everyone's problem (*tu quoque*) Probability-raising accounts get them backward generally, since Pam is a preemptor and Bob a fizzler Process-linkage accounts get it right that Bob is a noncause, but only because they cannot count Pam as a cause at all since she is a disconnector

Preemptive disconnections also illustrate the intransitivity and discontinuity of Analysis 1 Intransitivity Pam's saboteuring is a PROP of the destruction of the tower, which is a PROP of the crash, but as per above her saboteuring is not a PROP of the crash Discontinuity Bob's saboteuring is a PROP of the crash, but there is a discontinuity where he fizzles

I see two options First, one might bite the bullet on all three objections Transitivity has been questioned before,[28] continuity is equally questionable, and perhaps one could learn to live with a counterintuitive verdict in preemptive disconnections, especially as long as no alternative analysis of causation fares better

A second option, which I prefer, is to refine the analysis It turns out that all three objections can be handled, and the original solutions to preemptions, traces, disconnections, and fizzlings maintained, by adding a continuity requirement

So I propose

Analysis 2 C causes E if and only if C is a *continuous* PROP for E

The continuity in question must be with respect to something, and the obvious candidate is space-time Thus (at least in worlds for which the causal order is the temporal order), I further interpret Analysis 2 as

> *Analysis 2 Interpreted* C is a continuous PROP for E if and only if there is a chain of *direct* PROPs between C and E, where C is a direct PROP for E if and only if (a) C is a PROP for E, and (b) for all times t_D between t_C and t_E, there is a D at t_D such that C is a PROP for D, and D is a PROP for E

Analysis 2 answers the transitivity and continuity objections, while handling preemptive disconnections Transitivity is gained, because if C causes D and D causes E then it follows that there is a chain of direct PROPs between C and D, and between D and E, and therefore a chain of direct PROPs between C and E Continuity is gained explicitly And in the preemptive disconnection case above, Pam's saboteuring will count as a direct PROP of the blowing up of the tower, and the blowing up of the tower will count as a direct PROP of the crash, and so Pam will come out a cause Bob, on the other hand, cannot be connected to the crash by direct PROPs because there is a discontinuity involved at the point where he is overcome with laughter and fizzles out there is no event *D* to be found at that time such that both Bob's saboteuring is a PROP for *D* and *D* is a PROP for the crash (Bob's laughter fails the latter) [29]

---

[28] Hartry Field proposes the following counterexample to transitivity your enemy places the bomb by your door (C), which causes your friend to defuse the bomb (D), which causes you to survive to eat breakfast next morning (E) But surely your enemy placing the bomb by your door (C) does not cause you to survive to eat breakfast (E)! As the reader may confirm in this example, C is a PROP for D (as an essential part of the D-process), D is a PROP for E (as a shield of the E-process), but C is not a PROP for E For further proposed counterexamples to transitivity, see also Michael McDermott, "Redundant Causation," *British Journal for the Philosophy of Science*, XL (1995) 523-44, Hall (*op cit*) and "Causation and the Price of Transitivity," this JOURNAL, XCVII, 4 (April 2000) 198-222

As the reader may confirm, Analysis 2 preserves the solutions to preemptions, fizzlings, disconnections, and traces The preemptor will count as itself a direct PROP (no chaining needed) of the effect, with the continuous path of the preempting process providing the relevant continuously present *D* events The fizzler will not count as a direct PROP of the effect and no chaining of direct PROPs will rebuild a connection, since the fizzle point will always constitute a discontinuity The disconnector will count as a direct PROP, with the continuous *D* events being the sequence of events of the destruction of the tower, followed by the sequence of absences of the alert transmission to the pilot, followed by the sequence of the pilot maintaining course And the trace will not count as a direct PROP and no chaining of direct PROPs will rebuild a connection, since though there is a continuous PROP sequence of the trace arriving at the effect, the irrelevance of any part of that sequence to the end always remains

But there is a price to be paid it looks as if a continuity requirement must stipulate away the seeming possibility of spatiotemporally discontinuous causations [30] If Merlin casts a "prince to frog" spell at noon and the prince turns into a frog at midnight, and there is a law directly connecting magic spells at noon to enchantments at midnight, then it seems that Merlin's spell is a discontinuous cause of the enfrogging Analysis 1 has no problem with such cases the process involved will be the two-membered sequence ⟨spell, enchantment⟩, and the spell will be an essential part of such a sequence But Analysis 2 will find no event *D* at, for example, 6 00 pm such that Merlin's spell is a PROP for *D*, and *D* is a PROP for the enfrogging

---

[29] If we merely take the transitive closure of Analysis 1, this will fail to exclude Bob

[30] If there is some independent way of defining the "causal dimension," then the PROPs relation can be required to be continuous with respect to that dimension rather than space-time This would solve the problem, but I do not know how to define such a dimension

I see three options retreat to Analysis 1, bite the bullet and declare discontinuous causation impossible,[31] or refine the analysis further to maintain the original solutions, handle preemptive disconnections, and allow discontinuous causation

The causally relevant difference between Bob the fizzled shield and Merlin the successful spellcaster seems to be that, although both are PROPs for their respective *E*s, there is a causal gap in the line of PROPs from Bob's saboteuring to the collision, and no causal gap in the PROP relation from Merlin's spellcasting to the enchantment If one can determine which PROP relations need intermediaries and which do not (obviously without relying on causal or temporal gaps) then one can require continuity with respect to needed intermediaries only, rather than with respect to space-time, and thereby achieve a successful further refinement

But I do not know how to determine when intermediaries are needed Some not implausible suggestions are available perhaps one can study whether *C* remains a PROP for *E* on the counterfactual supposition of certain candidate needed-intermediaries not occurring, or perhaps one can study whether there are intermediaries *D* (where *C* is a PROP for *D* and *D* is a PROP for *E*) in certain unproblematic exemplars of such a *C* causing such an *E* I do not know whether these suggestions will work generally More work remains to be done

The nature of the causal relation remains mysterious, but I believe that in important respects both Analysis 1 and Analysis 2 shed more light on this relation than any extant alternatives, and that there may be some Analysis 3 nearby that will yield complete success I conclude that causes are best understood as probability raisers of processes, though exactly how best the story goes remains an open question

JONATHAN SCHAFFER

University of Massachusetts/Amherst

---

[31] Mellor (*op cit*) regards spatiotemporal continuity as a central connotation of causation, and the other successes of Analysis 2 may be taken as an indirect argument for Mellor's claim

BOOK REVIEWS

Responsibility and Control A Theory of Moral Responsibility JOHN MARTIN FISCHER and MARK RAVIZZA New York Cambridge University Press, 1998 vii + 277 p Cloth $64 95, paper $19 95 *

In *Responsibility and Control A Theory of Moral Responsibility,* John Martin Fischer and Mark Ravizza defend a distinctive compatibilist account of the control (or free will) condition necessary for moral responsibility They maintain that this condition requires guidance and not regulative control Guidance control consists in the power to perform an action freely, regulative control consists in the power to perform an action freely, and also the power to act otherwise (31) Their analysis of guidance control yields a reasons-responsive, actual-sequence, mechanism-based model Reason responsiveness characterizes the pertinent control in terms of the capacity to respond appropriately to (at least some) reasons for action (41-46) An actual-sequence account illuminates control purely in terms of the (dispositional) properties present in the actual causal sequence of events giving rise to freely willed action, and not even partially in terms of the properties manifest in alternative courses of action (37-38) A mechanism-based account does *not* attribute control to an agent, but to more carefully specified features directly involved in the etiology of an agent's action—features that, Fischer and Ravizza explain, must be the agent's own in an important sense (38-39)

Mechanism-based accounts as opposed to agent-based accounts give reasons-responsive, actual-sequence compatibilist models a crucial advantage in explaining the guidance control at work in an example of the sort made famous by Harry Frankfurt These controversial examples attempt to show, typically in defense of compatibilism, cases in which a paradigmatically morally responsible agent acts freely but in which circumstances *playing no role in an agent's freely willed action* nevertheless ensure that the agent cannot do other than she does Notice that, in such cases, an *agent* is not reasons responsive (38) Due to the ensuring conditions, were the agent presented with reason to act otherwise, the agent would not be able to do so But

---

* I would like to thank Bernard Berofsky, Carl Ginet, Ishtiyaque Haji, Al Mele, Derk Pereboom, Eleonore Stump, and David Widerker for excellent comments on earlier drafts of this review I owe a special debt of gratitude to John Martin Fischer for his thorough comments and generous support

0022-362X/01/9802/93–100

focus instead upon the agent's mechanism of action in the actual sequence of events If the *mechanism* were operative, and if different reasons were presented to the agent, then the agent, by way of this mechanism, would do otherwise (38-39) (Of course, in worlds in which such a mechanism is operative and the agent has reason to do otherwise, the restrictive ensuring conditions are not in place) Hence, a mechanism-based approach permits Fischer and Ravizza a reasons-responsive, actual-sequence analysis of control even in Frankfurt cases This is an enormous theoretical advantage

I

There is a great deal at stake in Fischer and Ravizza's work Theirs is the most advanced, thorough, and provocative treatment of compatibilism to date Several features of their account make this so Here are four

First, by offering a convincing account of guidance control—an account that does not require the ability to do otherwise—Fischer and Ravizza are able to avoid the most powerful case for incompatibilism that determinism rules out the freedom to do otherwise, and the inability to do otherwise rules out moral responsibility (33-34)

Second, Fischer and Ravizza's account elegantly advances an insight about reasons responsiveness central to Susan Wolf's highly influential work in, among other places, *Freedom within Reason*[1] Wolf's view, while enticing, leaves the notion of reasons responsiveness unanalyzed It is also open to incompatibilist worries about blameworthiness Why? For Wolf, praiseworthiness and blameworthiness are asymmetrical in their control conditions (*ibid*, pp 79-81) While praiseworthiness requires only guidance control, blameworthiness requires regulative control But this reliance on regulative control invites classic incompatibilist worries about the ability to do otherwise Fischer and Ravizza's account remedy both of these shortcomings by carefully articulating the conditions for reasons responsiveness (62-91), and also by showing that the intuitions driving Wolf to treat moral responsibility for the praiseworthy as resting upon guidance control (and not alternative possibilities or regulative control) can just as easily be brought to bear upon the blameworthy (55-61)

Third, Fischer and Ravizza's compatibilism can be understood as a defense of the compatibilist thesis advanced by Daniel Dennett in *Elbow Room Varieties of Free Will Worth Wanting*[2] Dennett, too, maintained that the freedom associated with alternative (and not actual) causal sequences is otiose (*ibid.*, p. 132). Fischer and Ravizza, however, take far more seriously than Dennett was willing to, incompatibilist arguments for the incompatibility of determinism and the freedom to do otherwise (17-25, and 33-34). Acknowledging that such freedom *would* be sufficient to satisfy the control necessary for moral responsibility, Fischer and Ravizza concede that this kind of freedom *is* incompatible with determinism. Also, in *Elbow Room* Dennett swiftly dismissed Frankfurt's famous compatibilist argument based on counterexamples. Dennett stated that such examples, while interesting, were "insufficiently ambitious" (*ibid.*, p. 132). As such, they could not bear the fruit of a general account of compatibilist freedom. Fischer and Ravizza, alternatively, make use of the Frankfurt examples as a key ingredient in their demonstration that it is only guidance control that is necessary for moral responsibility. On this score, while Dennett simply missed the boat, Fischer and Ravizza built a better one!

---

[1] New York Oxford, 1990

[2] Cambridge MIT, 1985

Fourth, Fischer and Ravizza's defense of compatibilism also makes innovative appeal to historical considerations as a condition for freedom and responsibility (170-206). Compatibilism has long been threatened by manipulation cases that show that all the requisite "compatibilist-friendly" control conditions could be satisfied by some, intuitively, responsibility-undermining process of manipulation. If so, then what makes determinism different from, say, a demon's process of manipulation? Fischer and Ravizza give an historical account of how an agent's reasons-responsive mechanism of action comes to be her own (207-39). They argue that such a process cannot, in principle, be the upshot of a process of manipulation, yet it could be realized in a determined world. This aspect of their account also rightly earns their work a prominent status within the literature.

For all of the above reasons, Fischer and Ravizza's *Responsibility and Control* represents an outstanding advance in the free-will debate. Despite these attributes, in what follows, by focusing upon their use of mechanism individuation, I shall argue that their defense of an actual-sequence model is threatened on plausible incompatibilist grounds.

II

In advancing their account of reasons responsiveness, Fischer and Ravizza devote considerable attention to a proper articulation of the degree of responsiveness required to capture the contours of morally responsible agency. They attempt to move between two extremes, one in which an agent acts from a reasons-responsive mechanism that is sensitive to any sufficient reason to do otherwise, another in which an agent acts from a reasons-responsive mechanism that is sensitive to perhaps just one reason to act otherwise (or instead some very small and perverse range of reasons to act otherwise) (63-68) One extreme sets standards for morally responsible agency too high, another too low To advance their view, Fischer and Ravizza make use of the notion of a "mechanism" issuing in action They thus unpack reasons responsiveness by first *holding fixed the mechanism* issuing in action in the actual world Then they consider possible worlds in which the "same" mechanism is active and in which different reasons to act otherwise are presented to the agent In some of those worlds, Fischer and Ravizza maintain, the agent does act otherwise This range of worlds, and of reasons, according to Fischer and Ravizza, must demonstrate a rational pattern (69-85) Their work here is subtle, and the account of "moderate reasons responsiveness" which emerges is compelling It hangs, however, on their ability to make use of the notion of "same mechanism "

Fischer and Ravizza openly admit that they have no principled basis for individuating mechanisms They simply settle for natural or intuitive judgments in various cases (47) As with sameness of smile or sameness of house, Fischer and Ravizza explain, sameness of mechanism does not require sameness right down to microdetails (52) Just as we can recognize a good friend's "same old smile" over the years, despite an aging mug, so too can we recognize "same mechanism issuing in action" despite variations in things like the reasons present to the agent in worlds in which the same mechanism operates [3]

Fischer and Ravizza state that by 'mechanism' they do not mean to point to "anything over and above the process that leads to the relevant upshot" (38) But, construed most naturally, the process that leads to an agent's action includes the entire complex of proximal antecedent events and states figuring into a correct explanation of why the action occurred [4] How should the salient features from this complex be isolated as those figuring in the particular mechanism issuing in action? If, despite Fischer and Ravizza's denial, sameness down to the microdetails were required for sameness of mechanism, their account of reasons responsiveness would face serious troubles. Suppose that the blameworthy reason (desire, intention, or belief) upon which an agent acts in the actual world is identical with (or supervenes upon) some neurophysiological state of the brain. If 'same mechanism' requires sameness of microdetails, then, when different sufficient reasons to do otherwise are present in another possible world, the operative mechanism in the possible world would have to respond to reasons which shared the same neurophysiological details as those associated with the reasons present to the agent in the actual world. But this would be nomically implausible (albeit not metaphysically impossible). If, as seems likely, the different reasons presented to the agent in the other possible world did not share the same microdetails, then in that possible world, *holding fixed the mechanism operative in the actual world*, the agent might not respond appropriately to the sufficient reason to do otherwise. Minimally, these reflections indicate that Fischer and Ravizza's thesis would be susceptible to refutation based upon likely empirical data.

---

[3] The only restrictions Fischer and Ravizza place on the mechanisms at issue is that they must be temporally intrinsic Thus, the temporally extrinsic mechanism *deliberation preceding the donation of five dollars to a charity* entails that an agent give five dollars to a charity in any world in which that mechanism operates As such, this mechanism is not the kind at issue with regard to guidance control (46-47) The temporally intrinsic mechanism, *deliberation involving a donation to a charity*, does not entail that an agent donate to a charity at a later time

[4] I thank Carl Ginet for this insight

It is therefore crucial to Fischer and Ravizza's approach that they exclude a narrow construal of sameness of mechanism. Yet because they offer no *principled* basis for mechanism individuation, they must rest their thesis purely on *intuitive* reactions to different cases. But, it might be objected, *which* elements from the entire complex (of proximal events and states antecedent to an action) should figure intuitively into the relevant mechanism will vary relative to explanatory perspective. The neurophysiologist's basis of parsimony will be different than that employed in everyday folk-psychological discourse. What reason have we to assume that Fischer and Ravizza's basis for individuation is the correct one?

The situation worsens if one pushes for a hyper-restricted notion of sameness of mechanism. On the hyper-restricted construal, the entire complex of proximal antecedent events and states function as the pertinent mechanism. If this were the relevant mechanism, an agent could not act from a reasons-responsive mechanism at a deterministic world. Any possible world in which the "same" mechanism operates at some time must share *all* the same facts with the actual world at that time. But if determinism is true in the actual world, there is no possible world which shares the same facts with the actual world at a time and in which, at a subsequent time, an agent acts differently in the possible world than she does in the actual world. On the hyper-restricted construal, only the incompatibilist could make sense of a reasons-responsive mechanism. Only at an indeterministic world could it be the case that the very mechanism operating in the actual world operates in a possible world and, on the basis of different reasons, the agent acts differently at that possible world.[5]

Some incompatibilists might be inclined to draw their line in the sand here, insisting upon a hyper-restricted notion of mechanism individuation. But, anticipating this move, Fischer and Ravizza reply that this would be question begging (52, note 22). The source of the incompatibilist's insistence would arise from the very thought that the appropriate responsiveness requires that the agent be able to act differently in the same world as the actual world in which she acts. This is precisely what is at stake. On this point, Fischer and Ravizza seem right. Yet it also seems problematic for Fischer and Ravizza to lean exclusively on intuitions about different cases in order to work with a looser notion of sameness of mechanism. Without any principled way to individuate mechanisms they are vulnerable to a similar charge—that their intuitions regarding sameness simply mask their compatibilist convictions that responsiveness *does not* require that an agent be able to act differently in the actual world.

There are other places where Fischer and Ravizza lean too heavily upon their intuitive judgments regarding sameness of mechanism. For instance, they distinguish two components of reasons responsiveness. *Reasons receptivity* allows agents to recognize and evaluate reasons for action (69-73). Here, Fischer and Ravizza place considerable constraints on the range of reasons to which a morally responsible agent's mechanism of action must be receptive. *Reasons reactivity* allows agents to act on the basis of the reasons for action (recognized at the receptivity stage) (73-76). Fischer and Ravizza permit a morally responsible agent's mechanism of action a quite limited and "idiosyncratic" pattern of reactivity (74-75). They hold that a morally responsible agent that is blameworthy for her action must act from a mechanism that would at least react to *some* reason to act otherwise, were some other reason to do so present in another possible world.

Given Fischer and Ravizza's account of reasons reactivity, here is a reasonable (even predictable) incompatibilist objection: reacting to a different sufficient reason to do otherwise in some other possible world will not show why an agent should be held blameworthy for *not* reacting to a morally sufficient reason to do otherwise in the actual world. Is not the relevant question, 'Could the agent have reacted differently to the very reason that was presented to her in the actual world?' Fischer and Ravizza respond by claiming that "reactivity is all of a piece " Reacting to a different sufficient reason to do otherwise in another possible world proves a general capacity to react to *any* sufficient reasons to do otherwise Hence they conclude that the agent's mechanism retains this general capacity in the actual world There is a great deal at stake in this move Without it, to show that a blameworthy agent acts from a reasons-reactive mechanism, one would have to show that there are possible worlds in which the agent reacts differently to the *same* moral reasons as those present in the actual world But, as explained above, this would involve incompatibilist-friendly regulative control

---

[5] At an indeterministic world, it need not be determined which of competing reasons the agent decides to act on, nor, for that matter, what reasons bear upon an agent's deliberations.

When defending their claim that reactivity is all of a piece the authors ask how it might be that reactivity is *not* all of a piece That is, how might it be that an agent is able to react to a sufficient reason to do otherwise in one possible world and yet does not retain the general capacity to react to different reasons in other possible worlds (in which the same mechanism operates)? They offer the following speculation

> How might this have turned out not to be true? Imagine that the agent somehow gets considerably more energy or focus if he is presented with a *strong* reason to do otherwise, and it is only in virtue of these factors that he successfully reacts to the reason (74)

Here is how Fischer and Ravizza respond

> There certainly can be cases like this, but it is natural to say that, when the agent acquires significantly more "energy or focus," this gives rise to a *different mechanism* from the actual mechanism Our point is that, *holding fixed the actual kind of mechanism,* reactivity is all of a piece if the mechanism can react to any reason to do otherwise, it can react to all such reasons (74)

Fischer and Ravizza cast the thought experiment in terms of the *agent's* getting more energy or focus when presented with strong reasons Of course, the question is, can the agent's *mechanism* gain more energy or focus when presented with stronger reasons? If it could, then it would not be the case that reactivity is "all of a piece " If one grants the intelligibility of talk of 'mechanisms' as opposed to agents in the first place, then it is hard to understand why it is any less natural to explain the thought experiment by saying that the very mechanism issuing in one action might get a different amount of charge or focus in the presence of some reasons than in the presence of others Because so much rides on their claim that reactivity is all of a piece, and because (granting talk of mechanisms) there are competitors as to what seems a natural way to explain the thought experiment, Fischer and Ravizza's appeal to what seems most natural looks out of place.

III

Fischer and Ravizza offer a compelling account of reasons-responsive, actual-sequence compatibilism. On their view, the best case for compatibilism proceeds by focusing upon the properties manifest in the actual sequence of events leading to action. More traditional compatibilist approaches foundered by worrying about alternative courses of action requiring access to alternative causal sequences. This invited forceful incompatibilist arguments (such as the well-known consequence argument) about the incompatibility of determinism and alternative possibilities. By offering a dispositional analysis of the properties present in the actual casual source of an agent's action, Fischer and Ravizza disarm this incompatibilist strategy. All they need to do is offer an analysis that captures an intuitively plausible basis for the control manifested by the agent in the *actual* world. Locating the pertinent control in an agent's reasons responsiveness (via an actual-sequence mechanism) certainly does that. This is a graceful philosophical maneuver, painstakingly worked out in clear and provocative detail. Their efforts have rightly placed their compatibilist view at the center of contemporary debate.

Fischer and Ravizza's appeal to sameness of mechanism is the lynchpin in their defense of an actual-sequence, reasons-responsive analysis of guidance control. Regrettably, their exclusive reliance on intuition as a basis for mechanism individuation renders their defense of their overall theory unconvincing. There are too many pressure points at which differing intuitions regarding sameness of mechanism yield troubling results for their defense of guidance control. Thus, to defend their compatibilist account of moral responsibility fully they must address this source of trouble.

MICHAEL MCKENNA

Bryn Mawr College
Ithaca College

MEMO TO AUTHORS

The *Journal of Philosophy* is published monthly, in issues that (ordinarily) contain only 56 pages of copy, this limits the length of papers the editors can normally publish to no more than 7,500 words, or eighteen *Journal* pages A more precise way of determining length is by characters 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages

The editors request that three copies of all manuscripts submitted be prepared according to the following guidelines

(1) All material, including the quotations and footnotes, must be typed, double-spaced, and one-sided, with ample margins left and right, top and bottom
(2) No bibliographies will be published, bibliographical material should be put into the footnotes No appendices will be published either
(3) Footnotes should be complete, according to *Journal* style (see any issue) Journal citations should include name of journal in full, volume number, issue number, month or season, year, and pages Book references should include publisher, U S city of publication, and date Footnotes should be gathered at the end of the article
(4) Quotations should be checked by the author, capitalization, italics, spelling, and the like should match the original
(5) If there are diagrams, they should be provided camera-ready by the author If there are special symbols, they should be widely spaced and clearly marked Standard logic symbols are available, and letters of the Greek and script alphabets
(6) No boldface or underlining, please limit use of italics to technical terms (when they are introduced) and to variables
(7) No contractions

The editors are willing to read and evaluate a manuscript that does not meet these guidelines, provided the author agrees to meet them before submitting a final manuscript The editors may ask for revision of excessively long manuscripts, and may require reformatting of accepted manuscripts They also request a disk with the final hard copy

The *Journal of Philosophy* does not publish papers that are primarily historical or expository, the editors urge that such manuscripts be submitted elsewhere While we do not publish manuscripts that mainly comment on articles published elsewhere, we do have a "Comments and Criticism" section in which short responses to *Journal* articles may be published

Manuscripts may be sent to Michael Kelly, Managing Editor, *Journal of Philosophy*, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, New York, NY 10027

NEW BOOKS

AMERIKS, KARL *Kant and the Fate of Autonomy Problems in the Appropriation of the Critical Philosophy* New York Cambridge, 2000 xiii, 351 p Cloth $54 95, paper $19 95

ANDERSON-GOLD, SHARON *Unnecessary Evil History and Moral Progress in the Philosophy of Immanuel Kant* Albany SUNY, 2001 xiii, 138 p Paper $16 95

ANGUS, IAN *(Dis)figurations Discourse/Critique/Ethics* New York Verso, 2000 xii, 269 p Paper $20 00

ANTONACCIO, MARIA *Picturing the Human The Moral Thought of Iris Murdoch* New York Oxford, 2000 vii, 244 p Cloth $35 00

ARGYLE, MICHAEL *Psychology and Religion An Introduction* New York Routledge, 2000 x, 285 p Paper $15 95

APPELBAUM, DAVID *The Delay of the Heart* Albany SUNY, 2000 xiii, 167 p Paper $19 95

BAKAR, OSMAN *The History and Philosophy of Islamic Science* Cambridge, UK Islamic Texts Society, 1999 x, 266 p Paper $25 00

BARKER, PHILIP *Michel Foucault An Introduction* Edinburgh University Press, 2000 xiii, 160 p Paper $19 00

BARNES, MICHAEL HORACE *Stages of Thought The Co-Evolution of Religious Thought and Science* New York Oxford, 2000 ix, 334 p Cloth $45 00

BARRETT, JEFFREY ALAN *The Quantum Mechanics of Minds and Worlds* New York Oxford, 1999 xv, 267 p Cloth $45 00

BAUDRILLARD, JEAN *The Vital Illusion* JULIA WITWER, ed New York Columbia, 2000 102 p Cloth $17 95

BENCIVENGA, ERMANNO *Hegel's Dialectical Logic* New York Oxford, 2000 xi, 143 p Cloth $32 00

BENSO, SILVIA *The Face of Things A Different Side of Ethics* Albany SUNY, 2000 xxxviii, 258 p Paper $18 95

BERLINSKI, DAVID *Newton's Gift How Sir Isaac Newton Unlocked the System of the World* New York Free Press, 2000 xviii, 217 p Cloth $24 00

BERNARDETE, SETH *The Argument of the Action* Chicago University Press, 2000 xxi, 434 p Cloth $39 00

BLATTBERG, CHARLES *From Pluralist to Patriotic Politics Putting Practice First* New York Oxford, 2000 xv, 294 p Cloth $85 00

BOHM, DAVID, and F DAVID PEAT *Science, Order, and Creativity,* Second Edition New York Routledge, 2000 viii, 316 p Paper $17 95

BORGMANN, ALBERT *Holding onto Reality The Nature of Information at the Turn of the Millennium* Chicago University Press, 1999 274 p Paper $14 00

BOURGEOIS, PATRICK L *Philosophy at the Boundary of Reason Ethics and Postmodernity* Albany SUNY, 2001 xxi, 280 p Paper $24 95

BRANDOM, ROBERT B *Articulating Reasons An Introduction to Inferentialism* Cambridge Harvard, 2000 230 p Cloth $35 00

BRANN, EVA *What, Then, Is Time?* Lanham, MD Rowman and Littlefield, 1999 xvi, 237 p Cloth $63 00, paper $23 95

BRENNAN, GEOFFREY, and ALAN HAMLIN. *Democratic Devices and Desires* New York· Cambridge, 2000 x, 267 p Cloth $59 95, paper $22 95

BRENNAN, TAD *Ethics and Epistemology in Sextus Empiricus* New York Garland, 1999 xiv, 115 p Cloth $45 00

BREWER, TALBOT *The Bounds of Choice· Unchosen Virtues, Unchosen Commitments* New York Garland, 2000 xiii, 237 p Cloth $55 00

BRILLIANT, RICHARD *My Laocoon Alternative Claims in the Interpretation of Artworks* Berkeley California UP, 2000 xvi, 146 p Cloth $45 00

BRINK, BERT VAN DEN *The Tragedy of Liberalism. An Alternative Defense of a Political Tradition* Albany SUNY, 2000 ix, 270 p Cloth $59 50, paper $19 95

BROOK, ANDREW, and ROBERT J STAINTON *Knowledge and Mind A Philosophical Introduction.* Cambridge MIT, 2000 xiv, 253 p. Cloth $27 95

BUCHANAN, ALLEN, DAN W BROCK, NORMAN DANIELS, and DANIEL WIKLER *From Chance to Choice Genetics and Justice.* New York Cambridge, 2000 xiv, 398 p Cloth $29 95

BUCHANAN, IAN *Deleuzism A Metacommentary* Durham Duke, 2000 vii, 209 p Cloth $49 95, paper $17 95

BURTNESS, JAMES H *Consequences Morality, Ethics, and the Future* Minneapolis Fortress, 1999 ix, 173 p Paper $16 00

BUTLER, JUDITH, ERNESTO LACLAU, and SLAVOJ ŽIŽEK *Contingency, Hegemony, Universality Contemporary Dialogues on the Left* New York Verso, 2000 vi, 329 p Paper $20 00

BYRNE, PETER *Philosophical and Ethical Problems in Mental Handicap* New York St Martin's, 2000 xiii, 175 p Cloth $59 95

CAPUTO, JOHN D *More Radical Hermeneutics On Not Knowing Who We Are* Bloomington Indiana UP, 2000 ix, 296 p Cloth $44 95, paper $19 95

CARE, NORMAN S *Decent People* New York Rowman and Littlefield, 2000 xv, 199 p Cloth $65 00, paper $18 95

CARSTAIRS-MCCARTHY, ANDREW *The Origins of Complex Language An Inquiry into the Evolutionary Beginnings of Sentences, Syllables, and Truth* New York Oxford, 1999 xi, 260 p Paper $19 95

CASSIRER, ERNST *The Logic of the Cultural Sciences Five Studies* New Haven Yale, 2000 xliii, 134 p Cloth $30 00, paper $15 00

CENTORE, F F *Two Views of Virtue Absolute Relativism and Relative Absolutism* Westport, CT Greenwood, 2000 216 p Cloth $65 00

CHOMSKY, NOAM *New Horizons in the Study of Language and Mind* New York Cambridge, 2000 xvii, 230 p Cloth $54 95, paper $19 95

CLACK, BRIAN R *An Introduction to Wittgenstein's Philosophy of Religion* Edinburgh University Press, 2000 xii, 137 p Paper $25 00

CLARK, AUSTEN *A Theory of Sentience* New York Oxford, 2000 xii, 288 p Cloth $45 00

CLAYTON, PHILIP *The Problem of God in Modern Thought* Grand Rapids, MI Eerdmans, 2000 xv, 516 p Cloth $40 00

COHEN, G A *If You're an Egalitarian, How Come You're So Rich?* Cambridge Harvard, 2000 xii, 233 p Cloth $35 00

CONNELL, RICHARD J *From Observables to Unobservables in Science and Philosophy* Lanham, MD UP of America, 2000 x, 236 p Cloth $64 00, paper $37 50

CUNNINGHAM, SUZANNE *What Is a Mind? An Integrative Introduction to the Philosophy of Mind* Indianapolis Hackett, 2000 xii, 274 p Paper $16 95

CZIKO, GARY *The Things We Do* Cambridge MIT, 2000 xi, 290 p Cloth $37 95

DENHAM A E *Metaphor and Moral Experience* New York Oxford, 2000 viii, 366 p Cloth $70 00

DEVINE, PHILIP E *Natural Law Ethics* Westport, CT Greenwood, 2000 xiii, 202 p Cloth $59 95

DODD, JULIAN *An Identity Theory of Truth* New York St Martin's, 2000 ix, 199 p Cloth $59 95

DOWE, PHIL *Physical Causation* New York Cambridge, 2000 ix, 224 p Cloth $54 95

DRYZEK, JOHN S *Deliberative Democracy and Beyond Liberals, Critics, Contestations* New York Oxford, 2000 ix, 195 p Cloth $29 95

DWORKIN, RONALD *Sovereign Virtue The Theory and Practice of Equality* Cambridge Harvard, 2000 511 p Cloth $35 00

EDEL, MAY, and ABRAHAM EDEL *Anthropology Ethics The Quest For Moral Understanding* New Brunswick, NJ Transaction, 2000 ix, 280 p Paper $24 95

FAIRBANKS, SANDRA JANE *Kantian Moral Theory and the Destruction of the Self* Boulder, CO Westview, 2000 x, 244 p Cloth $69 00

FERGUSON, HARVIE *Modernity and Subjectivity Body, Soul, Spirit* Charlottesville Virginia UP, 2000 x, 218 p Cloth $49 50, paper $17 50

FLETCHER, SCOTT *Education and Emancipation Theory and Practice in a New Constellation* New York Columbia, 2000 xii, 211 p Cloth $50 00, paper $23 95

FLIKSCHUH, KATRIN *Kant and Modern Political Philosophy* New York Cambridge, 2000 x, 216 p Cloth $54 95

FLYNN, JAMES R *How to Defend Human Ideals Substitutes for Objectivity* Lincoln Nebraska UP, 2000 ix, 212 p Cloth $40 00

FODOR, JERRY *The Mind Doesn't Work That Way The Scope and Limits of Computational Psychology* Cambridge MIT, 2000 xi, 126 p Cloth $22 95

FONTINELL, EUGENE *Self, God, and Immortality A Jamesian Investigation* New York Fordham, 2000 xxv, 297 p Cloth $39 95, paper $19 95

FORSTER, ECKART *Kant's Final Synthesis An Essay on the* Opus postumum Cambridge Harvard, 2000 xx, 207 p Cloth $35 00

FOSS, JEFFREY *Science and the Riddle of Consciousness A Solution* Norwell, MA Kluwer, 2000 xiii, 225 p Cloth $99 50

FOSTER, JOHN *The Nature of Perception* New York Oxford, 2000 viii, 289 p Cloth $55 00

FROMM, ERICH *The Art of Loving* New York Continuum, 2000 viii,130 p Cloth $18 95

GARDINER, MARK QUENTIN *Semantic Changes to Realism* Toronto University Press, 2000 x, 267 p Cloth $85 00

GAUS, GERALD F *Political Concepts and Political Theories* Boulder, CO Westview, 2000 xiv, 288 p Paper $20 00

GELVEN, MICHAEL *The Asking Mystery A Philosophical Inquiry* University Park Penn State UP, 2000 177 p Cloth $35 00, paper $18 95, *Truth and the Comedic Art* Albany SUNY, 2000 vii, 172 p Paper $16 95

GINTIS, HERBERT *Game Theory Evolving A Problem-Centered Introduction to Modeling Strategic Behavior* Princeton University Press, 2000 xxxv, 531 p Cloth $55 00, paper $29 95

GLAZEBROOK, TRISH *Heidegger's Philosophy of Science* New York Fordham, 2000 xii, 278 p Cloth $39 95, paper $19 95

GLOVER, JONATHAN *Humanity A Moral History of the Twentieth Century* New Haven Yale, 2000 xiv, 464 p Cloth $27 95

GOTZ, IGNACIO L *Manners and Violence* Westport, CT Praeger, 2000 xiii, 131 p Cloth $52 50

GRIFFIN, DAVID RAY *Religion and Scientific Naturalism Overcoming the Conflicts* Albany SUNY, 2000 xvii, 345 p Paper $25 95

GUEVARA, DANIEL *Kant's Theory of Moral Motivation* Boulder, CO Westview, 2000 ix, 158 p Cloth $60 00

GUNNARSSON, LOGI *Making Moral Sense Beyond Habermas and Gauthier* New York Cambridge, 2000 xi, 286 p Cloth $59 95

GUTTENPLAN, SAMUEL *Mind's Landscape An Introduction to the Philosophy of Mind* Malden, MA Blackwell, 2000 viii, 358 p Cloth $62 95, paper $26 95

HACOHEN, MALACHI HAIM *Karl Popper The Formative Years* New York Cambridge, 2000 xii, 610 p Cloth $54 95

HALLETT, GARTH L *A Middle Way to God* New York Oxford, 2000 viii, 162 p Cloth $35 00

HANFLING, OSWALD *Philosophy and Ordinary Language Bent and Genius of Our Tongue* New York Routledge, 2000 xi, 205 p Cloth $85 00

HANLEY, CATRIONA *The Role of Method in Thinking the Infinite* New York Rowman and Littlefield, 2000 xxii, 217 p Cloth $60 00, paper $22 95

HARDIN, RUSSELL *Liberalism, Constitutionalism, and Democracy* New York Oxford, 1999 xviii, 379 p Cloth $29 95

HARVEY, DAVID *Spaces of Hope* Berkeley California UP, 2000 x, 293 p Cloth $24 95

HARVEY, PETER *An Introduction to Buddhist Ethics* New York Cambridge, 2000 xx, 478 p Cloth $54 95, paper $19 95

HASS, ANDREW *Hegel and the Problem of Multiplicity* Evanston, IL Northwestern, 2000 xxxii, 355 p Paper $29 95

HAYWARD, CLARISSA *De-facing Power* New York Cambridge, 2000 ix, 214 p Cloth $54 95, paper $19 95

HEDLEY, DOUGLAS *Coleridge, Philosophy, and Religion Aids to Reflection and the Mirror of the Spirit* New York Cambridge, 2000 xiv, 330 p Cloth $64 95

HELM, PAUL *Faith with Reason* New York Oxford, 2000 xvi, 185 p Cloth $39 95

HENRY, PAGET *Caliban's Reason Introducing Afro-Caribbean Philosophy* New York Routledge, 2000 xiii, 304 p Paper $21 95

HILL, CLAIRE ORTIZ, and GUILLERMO E ROSADO HADDOCK *Husserl or Frege? Meaning, Objectivity, and Mathematics* Chicago Open Court, 2000 xiv, 315 p Cloth $39 95

HILL, THOMAS E , JR *Respect, Pluralism, and Justice Kantian Perspectives* New York Oxford, 2000 xi, 281 p Cloth $60 00, paper $19 95

HOBBS, ANGELA *Plato and the Hero Courage, Manliness and the Impersonal Good* New York Cambridge, 2000 xvii, 280 p Cloth $59 95

HOBSON, JOHN M *The State and International Relations* New York Cambridge, 2000 viii, 258 p Cloth $52 95, paper $18 95

HOCHSTRASSER, T J *Natural Law Theories in the Early Enlightenment* New York Cambridge, 2000 xiii, 246 p Cloth $54 95

HONG, NATHANIEL J , KATHRYN HONG, and REGINE PRENZEL-GUTHRIE *Cumulative Index to Kierkegaard's Writings The Works of Søren Kierkegaard* Princeton University Press, 2000 xxi, 560 p Cloth $79 50

HOOKWAY, CHRISTOPHER *Truth, Rationality, and Pragmatism Themes from Peirce* New York Oxford, 2000 vi, 313 p Cloth $45 00

HURSTHOUSE, ROSALIND *On Virtue Ethics* New York Oxford, 2000 ix, 275 p Cloth $35 00

ISER, WOLFGANG *The Range of Interpretation* New York Columbia, 2000 xv, 206 p Cloth $25 00

JOHNSON, RALPH H *Manifest Rationality A Pragmatic Theory of Argument* Mahwah, NJ Erlbaum, 2000 xiii, 391 p Cloth $79 95, paper $36 50

KEKES, JOHN *Pluralism in Philosophy Changing the Subject* Ithaca Cornell, 2000 ix, 225 p Cloth $32 50

KELLER, EVELYN FOX *The Century of the Gene* Cambridge Harvard, 2000 186 p Cloth $22 95

KLOSKO, GEORGE *Democratic Procedures and Liberal Consensus* New York Oxford, 2000 x, 271 p Cloth $45 00

KRELL, DAVID FARRELL *The Purest of Bastards Works of Mourning, Art, and Affirmation in the Thought of Jacques Derrida* University Park Penn State UP, 2000 xiii, 237 p Cloth $45 00, paper $17 95

KUTZ, CHRISTOPHER *Complicity Ethics and Law for a Collective Age* New York Cambridge, 2000 xii, 331 p Cloth $59 95

LANDRY, LORRAINE Y *Marx and the Postmodernism Debates* Westport, CT Praeger, 2000 xiii, 232 p Cloth $65 00

LANGE, MARC *Natural Laws in Scientific Practice* New York Oxford, 2000 xvi, 348 p Cloth $60 00

LEAR, JONATHAN *Happiness, Death, and the Remainder of Life* Cambridge Harvard, 2000 189 p Cloth $24 00

LEHRER, KEITH *Theory of Knowledge,* Second Edition Boulder, CO Westview, 2000 xv, 250 p Paper $24 00

LOCKE, JOHN *The Reasonableness of Christianity as Delivered in the Scriptures* JOHN C HIGGINS-BIDDLE, ed New York Oxford, 1999 cxxxix, 261 p Cloth $95 00

LOCKHART, TED *Moral Uncertainty and Its Consequences* New York Oxford, 2000 xvi, 211 p Cloth $45 00

LORAND, RUTH *Aesthetic Order* New York Routledge, 2000 xi, 324 p Cloth $85 00

LOW, DOUGLAS *Merleau-Ponty's Last Vision A Proposal for the Completion of* The Visible and the Invisible Evanston, IL Northwestern, 2000 xv, 124 p Cloth $75 00, paper $19 95

LYCAN, WILLIAM G *Philosophy of Language A Contemporary Introduction* New York Routledge, 1999 xv, 243 p Cloth $65 00, paper $19 99

MACKEY, JAMES P *The Critique of Theological Reason* New York Cambridge, 2000 v, 333 p Cloth $59 95

MALLOY, ROBIN PAUL *Law and Market Economy Reinterpreting the Values of Law and Economics* New York Cambridge, 2000 x, 179 p Cloth $64 95, paper $22 95

MARTIN, MICHAEL *The Uses of Understanding in Social Science Verstehen* New Brunswick, NJ Transaction, 2000 x, 264 p Cloth $34 95

MASON, ANDREW *Community, Solidarity, and Belonging Levels of Community and Their Normative Significance* New York Cambridge, 2000 viii, 246 p Cloth $59 95, paper $21 95

MATILAL, BIMAL KRISHNA *The Character of Logic in India* JONARDON GANERI and HEERAMAN TIWARI, eds New York Oxford, 2000 ix, 180 p Cloth $22 95

MCCABE, MARY MARGARET *Plato and His Predecessors The Dramatisation of Reason* New York Cambridge, 2000 viii, 318 p Cloth $59 95

MCCLARY, SUSAN *Conventional Wisdom The Content of Musical Form* Berkeley California UP, 2000 xiv, 205 p Cloth $24 95

MCCUMBER, JOHN *Philosophy and Freedom Derrida, Rorty, Habermas, Foucault* Bloomington Indiana UP, 2000 xi, 191 p Cloth $39 95, paper $18 95

MCGHEE, MICHAEL *Transformations of Mind Philosophy as Spiritual Practice* New York Cambridge, 2000 viii, 293 p Cloth $59 95, paper $22 95

MEYER, LEONARD B *The Spheres of Music A Gathering of Essays* Chicago University Press, 2000 x, 316 p Cloth $55 00, paper $22 00

MILLER, PETER N *Peiresc's Europe Learning and Virtue in the Seventeenth Century* New Haven Yale, 2000 xvi, 234 p Cloth $40 00

MILLIKAN, RUTH GARRETT *On Clear and Confused Ideas An Essay about Substance Concepts* New York Cambridge, 2000 xiii, 258 p Cloth $59 95, paper $21 95

MIŠČEVIĆ, NENAD *Rationality and Cognition Against Relativism-Pragmatism* Toronto University Press, 2000 x, 325 p Cloth $95 00

MODRAK, DEBORAH K *Aristotle's Theory of Language and Meaning* New York Cambridge, 2000 ix, 302 p Cloth $59 95

MORGAN, DIANE *Kant Trouble The Obscurities of the Enlightened* New York Routledge, 2000 xii, 238 p Cloth $75 00, paper $24 99

MULLARKEY, JOHN *Bergson and Philosophy* Notre Dame University Press, 2000 xi, 206 p Paper $20 00

MURPHY, LIAM B *Moral Demands in Nonideal Theory* New York Oxford, 2000 ix, 168 p Cloth $35 00

MYERS, ROBERT H *Self-Governance and Cooperation* New York Oxford, 1999 179 p Cloth $45 00

NANDERSON, DESMOND *Songs without Music Aesthetic Dimensions of Law and Justice* Berkeley California UP, 2000 xiii, 303 p Cloth $55 00

NEUHOUSER, FREDERICK *Foundations of Hegel's Social Theory Actualizing Freedom* Cambridge Harvard, 2000 xiii, 337 p Cloth $49 95

NEWEY, GLEN *Virtue, Reason, and Toleration The Place of Toleration in Ethical and Political Philosophy* Edinburgh University Press, 1999 ix, 208 p Cloth $65 00

NIINILUOTO, ILKKA *Critical Scientific Realism* New York Oxford, 1999 xiv, 341 p Cloth $55 00

NORRIS, CHRISTOPHER *Minding the Gap Epistemology and Philosophy of Science in the Two Traditions* Amherst Massachusetts UP, 2000 xiii, 296 p Cloth $39 95, *Quantum Theory and the Flight from Realism Philosophical Responses to Quantum Mechanics* New York Routledge, 2000 ix, 266 p Cloth $85 00, paper $25 99

NOZICK, ROBERT *Teleology and the Norms of Nature* New York Garland, 2000 ix, 390 p Cloth $76 00

O'CONNOR, TIMOTHY *Persons and Causes The Metaphysics of Free Will* New York Oxford, 2000 xv, 135 p Cloth $35 00

O'SHAUGHNESSY, BRIAN *Consciousness and the World* New York Oxford, 2000 viii, 705 p Cloth $55 00

OWEN, DAVID *Hume's Reason* New York Oxford, 1999 x, 234 p Cloth $45 00

OWENS, DAVID *Reason without Freedom The Problem of Epistemic Normativity* New York Routledge, 2000 viii, 199 Paper $29 99

OYAMA, SUSAN *Evolution's Eye A Systems View of the Biology-Culture Divide* Durham Duke, 2000 vii, 274 p Cloth $54 95, paper $18 95, *The Ontogeny of Information,* Revised Edition Durham Duke, 2000 xxi, 273 p Cloth $59 95, paper $19 95

PAINE, THOMAS *Political Writings,* Revised Edition BRUCE KUKLICK, ed New York Cambridge, 2000 xxx, 347 p Cloth $42 95, paper $12 95

PAPPAS, GEORGE S *Berkeley's Thought* Ithaca Cornell, 2000 xi, 261 p Cloth $39 95

PAREKH, BHIKHU *Rethinking Multiculturalism Cultural Diversity and Political Theory* Cambridge Harvard, 2000 xii, 379 p Cloth $35 00

PENELHUM, TERENCE *Themes in Hume The Self, the Will, Religion* New York Oxford, 2000 xix, 294 p Cloth $55 00

PETTERSSON, ANDERS *Verbal Art A Philosophy of Literature and Literary Experience* Montreal McGill-Queens, 2000 x, 373 p Cloth $65 00

PIETROSKI, PAUL M *Causing Actions* New York Oxford, 2000 ix, 274 p Cloth $45 00

PILLOW, KIRK *Sublime Understanding Aesthetic Reflection in Kant and Hegel* Cambridge MIT, 2000 viii, 377 p Cloth $37 95

PORTER, JAMES I *The Invention of Dionysus An Essay on The Birth of Tragedy* Stanford University Press, 2000 xiii, 224 p Cloth $45 00, paper $16 95

POTTER, MICHAEL *Reason's Nearest Kin Philosophies of Arithmetic from Kant to Carnap* New York Oxford, 2000 x, 305 p Cloth $45 00

PUST, JOEL *Intuitions as Evidence* New York Garland, 2000 xiv, 135 p Cloth $45 00

RAWLS, JOHN *Lectures on the History of Moral Philosophy* BARBARA HERMAN, ed Cambridge Harvard, 2000 xxii, 384 p Cloth $45 00, paper $19 95

RAZ, JOSEPH *Engaging Reason On the Theory of Value and Action* New York Oxford, 1999 336 p Cloth $45 00

RECANTI, FRANÇOIS *Oratio Obliqua, Oratio Recta An Essay on Metarepresentation* Cambridge MIT, 2000 xviii, 360 p Cloth $65 00, paper $24 95

RESCHER, NICHOLAS *Inquiry Dynamics* New Brunswick, NJ Transaction, 2000 vii, 107 p Cloth $29 95

RICE, HUGH *God and Goodness* New York Oxford, 2000 xiv, 147 p Cloth $29 95

RICHTER, GERHARD *Walter Benjamin and the Corpus of Autobiography* Detroit Wayne State, 2000 309 p Cloth $29 95

RILEY, DENISE *The Words of Selves Identification, Solidarity, Irony* Stanford University Press, 2000 219 p Cloth $49 50, paper $18 95

ROSEN, CHARLES *Critical Entertainments Music Old and New* Cambridge Harvard, 2000 328 p Cloth $29 95

ROSENBERG, ALEXANDER *Darwinism in Philosophy, Social Science, and Policy* New York Cambridge, 2000 ix, 257 p Cloth $59 95, paper $19 95

ROSHWALD, MORDECAI *Liberty Its Meaning and Scope* Westport, CT Greenwood, 2000 204 p Cloth $65 00

RUBIO-MARÍN, RUTH *Immigration as a Democratic Challenge Citizenship and Inclusion in Germany and the United States* New York Cambridge, 2000 viii, 270 p Cloth $57 95, paper $21 95

RUSE, MICHAEL *Can a Darwinian be a Christian?* New York Cambridge, 2001 xi, 272 p Cloth $24 95

SALLIS, JOHN *Force of Imagination The Sense of the Elemental* Bloomington Indiana UP, 2000 xii, 237 p Cloth $49 95, paper $24 95

SASSOWER, RAPHAEL *A Sanctuary of Their Own Intellectual Refugees in the Academy* New York Rowman and Littlefield, 2000 xxi, 113 p Cloth $60 00, paper $22 95

SAVEDOFF, BARBARA E *Transforming Images How Photography Complicates the Picture* Ithaca Cornell, 2000 xi, 233 p Cloth $35 00

SAYERS, SEAN *Plato's Republic An Introduction* Edinburgh University Press, 2000 x, 178 p Paper $18 00

SCHÖNFELD, MARTIN *The Philosophy of the Young Kant* New York Oxford, 2000 xv, 348 p Cloth $55 00

SCHOULS, PETER A *Descartes and the Possibility of Science* Ithaca Cornell, 2000 xi, 171 p Cloth $35 00

SECADA, JORGE *Cartesian Metaphysics The Late Scholastic Origins of Modern Philosophy* New York Cambridge, 2000 xii, 333 p Cloth $59 95

SEGAL, GABRIEL M A *A Slim Book about Narrow Content* Cambridge MIT, 2000 ix, 177 p Cloth $50 00, paper $22 00

SHARPE, R A *Music and Humanism An Essay in the Aesthetics of Music* New York Oxford, 2000 ix, 221 p Cloth $35 00

SINGER, IRVING *George Santayana, Literary Philosopher* New Haven Yale, 2000 xiii, 217 p Cloth $25 00

SKLAR, LAWRENCE *Theory and Truth Philosophical Critique within Foundational Science* New York Oxford, 2000 x, 153 p Cloth $29 95

SKORUPSKI, JOHN *Ethical Explorations* New York Oxford, 1999 viii, 300 p Cloth $45 00

SMILANSKY, SAUL *Free Will and Illusion* New York Oxford, 2000 xii, 329 p Cloth $65 00

SMITH, JOHN H *Freedom, Power, and Understanding in Modern French and German Thought* Detroit Wayne State, 2000 239 p Cloth $39 95

SORABJI, RICHARD *Emotion and Peace of Mind* New York Oxford, 2000 x, 499 p Cloth $45 00

STEWART, JON *The Unity of Hegel's* Phenomenology of Spirit *A Systematic Interpretation* Evanston, IL Northwestern, 2000 xv, 556 p Cloth $69 95

STROLL, AVRUM *Twentieth-Century Analytic Philosophy* New York Columbia, 2000 ix, 302 p Cloth $32 50

STROUD, BARRY *Understanding Human Knowledge Philosophical Essays* New York Oxford, 2000 xix, 246 p Cloth $35 00, *Meaning, Understanding, and Practice Philosophical Essays* New York Oxford, 2000 xvii, 234 p Cloth $35 00

THAYER-BACON, BARBARA J *Transforming Critical Thinking Thinking Constructively* New York Columbia, 2000 xvi, 199 p Cloth $50 00, paper $23 95

TRAVIS, CHARLES *Unshadowed Thought* Cambridge Harvard, 2000 xiv, 263 p Cloth $39 95

TYE, MICHAEL *Consciousness, Color, and Content* Cambridge MIT, 2000 xiii, 198 p Cloth $29 95

WARDY, ROBERT *Aristotle in China Language, Categories, and Translation* New York Cambridge, 2000 x, 170 p Cloth $54 95

WEDIN, MICHAEL V *Aristotle's Theory of Substance The* Categories *and* Metaphysics *Zeta* xiii, 482 p Cloth $55 00

WHEELER, SAMUEL C , III *Deconstruction as Analytic Philosophy* viii, 294 p Cloth $49 50, paper $19 95

WHITE, STEPHEN K *Sustaining Affirmation* Princeton University Press, 2000 xii, 158 p Paper $15 95

WILLIAMS, JAMES *Lyotard and the Political* New York Routledge, 2000 ix, 153 p Cloth $65 00, paper $20 99

YAFFE, GIDEON *Liberty Worth the Name Locke on Free Agency* Princeton University Press, 2000 xii, 176 p Cloth $49 50, paper $16 95

YOLTON, JOHN W *Realism and Appearances* New York Cambridge, 2000 xiii, 157 p Cloth $54 95, paper $19 95

YOUNG, IRIS M *Inclusion and Democracy* New York Oxford, 2000 x, 304 p Cloth $29 95

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO 3, MARCH 2001

DISJUNCTIVE PROPERTIES MULTIPLE REALIZATIONS*

There has been much debate concerning the status of "disjunctive properties," and much of the debate is fueled by attacks and defenses of *nonreductive physicalism* (NRP) The most prevalent argument in support of NRP, the multiple realizability argument, is standardly thought to stand or fall depending upon the legitimacy of such properties John Heil[1] presents this standard line of thought as follows

> We unhesitatingly ascribe mental states across species, despite large differences in underlying physiology Even within our own species, it seems unlikely that particular mental characteristics are invariably realized in identical neural structures
>
> Multiple realizability, however, need not deter a determined identity theorist [that is, reductionist] It is open for such a theorist, for instance, to argue that the relevant characteristic is, in fact, disjunctive in character That is, it might be that, in you, mental feature M is realized in neural structure N, whereas in an octopus, M is realized in a different sort of neural structure N' Would this undermine type identity? It would not, unless we assume that M [could] not be identical with the disjunctive characteristic <N or N'> (*ibid*, p 64)

If there were legitimate disjunctive properties, the argument from multiple realizability would not refute reductionism Hence advo-

---

* Most of the ideas for this paper developed from John Heil's National Endowment for Humanities Summer Seminar, "Metaphysics of Mind," held at Cornell University in 1996 The paper owes much to Heil and all the participants, though Michael Watkins, Heather Gert, David Pitt, and Eric Saidel deserve special mention They are not, however, responsible for the faults in this paper The paper has also benefited from useful exchanges with Sydney Shoemaker, Alex Byrne, Sylvain Bromberger, Jeff McConnell, and the members of the philosophy departments at Carleton University and the University of Wisconsin/Madison

[1] *The Nature of True Minds* (New York Cambridge, 1992)

0022-362X/01/9803/111-36

cates of NRP argue that there are no disjunctive properties, or that they are in some way illegitimate [2] Contrary to this standard line of thought, I shall argue that there is good reason to grant the existence and legitimacy of disjunctive properties, and I shall sketch an account of such properties Yet I shall also argue that the existence and legitimacy of disjunctive properties is compatible with a robust, anti-eliminativist, version of NRP, a version of NRP that ought to satisfy psychologists and philosophers of mind (and economists, biologists, and the like)

I NRP, THE ARGUMENT FROM MULTIPLE REALIZABILITY, AND THE DISJUNCTIVE STRATEGY

NRP is defined as the conjunction of two theses

> *Physicalism* all particulars are constituted by physical particulars, and all properties are realized by physical properties [3]
>
> *Nonreducibility* mentalistic predicates cannot be reduced by physicalistic predicates

Before the argument from multiple realizability in support of NRP can be presented, some clarifying remarks concerning these theses are in order To clarify the thesis of physicalism, one must specify what it is for particulars (be they objects or events) and properties to be *physical*, and one must define the relations of *constitution* and *realization* The issue of what it is for particulars and properties to be physical is central and all too often overlooked, but I shall not here address it I shall also have nothing to say concerning the constitution relation, as my focus will be on properties and not particulars I shall, however, make a significant proposal concerning the realization relation, but for now a rough working definition of the *realization relation* will suffice a property *P* of an object (or event) *o* *realizes* a property *F* of *o* if and only if (i) it is necessary that, if *o* instantiates *P*, then *o* instantiates *F*, and (ii) *o*'s instantiating *P* in some metaphysical sense explains o's instantiating F—being P is one *way in which* a thing can be F So, for example, the property of being 85° C which is instantiated by my coffee cup is, it is plausible to suppose, realized by some very discriminating microphysical property which is also instantiated by my cup Given that my cup has that discriminating microphysical property, it is necessary that it also have the property of being 85° C And that my cup has that very discriminating microphysical property explains that my cup has the property of being 85° C—that particular microphysical property that my cup instantiates is one of many ways in which my cup could be 85° C Thus to say that all mental properties are *realized* by physical properties is to say that for every instantiation of a mental property M by an object (or event) o, there is some physical property P instantiated by o such that it is necessary that, if o instantiates P, then o instantiates M, and o's being P explains—in the robust metaphysical sense—o's being M[4]

---

[2] Influential defenders of NRP and critics of disjunctive properties include Hillary Putnam, "Psychological Predicates," in W H Capitan and D D Merrill, eds , *Art, Mind, and Religion* (Pittsburgh University Press, 1967), Ned Block and Jerry Fodor, "What Psychological States Are Not," *Philosophical Review*, LXXXI, (1972) 159-81, and Fodor, "Special Sciences," in *Representations* (Cambridge MIT, 1982), pp 127-45

[3] Two points First, a weaker version of *physicalism* requires only that all property instantiations be instantiated in physical particulars I think this weaker version of physicalism fails to capture the essence of NRP, but none of my arguments depends upon the stronger thesis And second, the theoretical role played by the notion of *realization* in this characterization of physicalism is often played by the notion(s) of *supervenience* It will become clear in section IV why I prefer to define physicalism in terms of realization

As I have formulated the thesis of nonreducibility, it concerns the nonreducibility of mentalistic *predicates*, and not the nonreducibility of mental *properties* In fact, strictly speaking the *reduction* relation holds between *theories*, but, at least within the framework of Ernest Nagel's[5] classic account of reduction, a derivative notion of reduction that holds between predicates can be specified On Nagel's account, very roughly, a theory T_1 is reduced to a theory T_2 just in case the laws that make up T_1 are shown to be derivable from the laws that make up T_2, together with a set of "bridge principles" that "connect" the predicates of T_1 with predicates of T_2 We may take the existence and truth of such a bridge principle that connects predicates ϕ and ψ to be necessary and sufficient for the predicate reduction of ϕ by ψ[6] Thus, the derivative notion of *predicate reduction* is defined as follows predicate ϕ is reduced by predicate ψ if and only if there is a universally quantified bridge principle that connects ϕ and ψ But now, of course, we must clarify what it is for there to be a universally quantified bridge principle that "connects" ϕ and ψ What relation must obtain between predicates ϕ and ψ if ϕ is to be connected to, and thereby reduced by, ψ?

---

[4] This sketch of the realization relation is essentially the account of realization proffered by Ernest LePore and Barry Loewer in "More on Making Mind Matter," *Philosophical Topics*, xvii, 1 (1989) 175-91

[5] See *The Structure of Science* (New York Harcourt, Brace and World, 1961) Nagel's account of theory reduction is sometimes criticized as requiring too much for theory reduction Thus, even if the defender of NRP is able to show that mentalistic predicates are not Nagel-reducible to physicalistic predicates, she has not thereby adequately defended NRP, for a weaker, more accurate, account of theory reduction may allow for the reducibility of the mental to the physical See, for example, John Bickle, "Multiple Realizability and Psychophysical Reduction," *Behavior and Philosophy*, xx, 1 (1992) 47-58 See also Patricia Churchland, *Neurophilosophy* (Cambridge MIT, 1986)

[6] On this account, the relation of *predicate reduction* is symmetric This is not problematic, as the relation of theory reduction is (still) asymmetric

The classic reductionists, or "type-type identity theorists," such as U T Place and J J C. Smart,[7] make it abundantly clear that a primary motivation for their theory is the ontological parsimony that results from identifying mental states with physical states In the sense in which *reduction* is relevant to the mind-body problem and the tenability of NRP, the reduction of one theory T_1 to another theory T_2 collapses the ontological commitments of T_1 to those of T_2, if T_1 is reduced to T_2, then one who endorses both theories is committed to the existence of only the entities posited by T_2 Consequently, the bridge principles by which a reduction is obtained must be interpreted as entailing, or at least warranting the assertion of, property identities (At least one must assent to this if one is to be a realist about properties and maintain that some predicates pick out, or designate, properties) So in asking, "What relation must obtain between predicates ϕ and ψ if ϕ is to be reduced by ψ?" we are asking, "What relation must obtain between two predicates if they are to be taken to designate the same property?"

Fortunately, my purposes here do not require me to provide an answer to the contentious question of what conditions are necessary and jointly sufficient for predicate reduction In order to clarify the argument from multiple realizability in support of NRP, and the "disjunctive strategy" in response to this argument, I need only the noncontentious assumption that a *necessary* condition for the predicate reduction of ϕ by ψ is that a universally quantified biconditional of the form $\ulcorner \forall x\, (\phi x \leftrightarrow \psi x) \urcorner$ be *nomologically necessary* Whatever might be *sufficient* for predicates ϕ and ψ to designate the same property, it is at least *necessary* that such a universally quantified biconditional hold with nomological necessity

The argument from multiple realizability is really an argument only in support of the thesis of nonreducibility, indeed, the argument is sometimes formulated in such a way that it presupposes the thesis of physicalism The argument was first sketched by Putnam (*op cit*) and then clarified by Fodor (*op cit*) and it has been widely discussed since Note that the argument from multiple realizability is alleged to demonstrate that the nonreducibility of mentalistic predicates by physicalistic predicates follows from the fact that mental properties are multiply realized by physical properties It is not obvious, however, how nonreducibility, which is at least in part an epistemological claim concerning *predicates*, could follow from the metaphysical claim that mental *properties* are multiply realized by physical properties A mental property M is multiply realized by physical properties if and only if it has some *realization base* of physical properties P_1, P_2, P_n ($n > 1$) where for each P_i in the realization base, P_i realizes M[8] Suppose that mental property M is multiply realized by properties P_1, P_2, P_n, and that a predicate μ designates M, and finally that for each P_i there is a corresponding predicate π_i designating P_i Further suppose that P_1, P_2, P_n are all *metaphysically incompatible* with each other, that is, for all objects o, o cannot instantiate both P_1 and P_2, nor both P_1 and P_3, and so on It follows that M is not identical to any of the P_i[9] Moreover, assuming that μ is predicate reduced by π_i only if a biconditional bridge principle of the form $\ulcorner\forall x\ (\mu x \leftrightarrow \pi_i x)\urcorner$ is (at least) nomologically necessary, it follows that μ is not predicate reduced by any π_i each P_i in M's realization base is sufficient, but not necessary, for M Consequently, each π_i is such that satisfying π_i is sufficient, but not necessary, for satisfying μ, and thus no bridge principle with the form displayed above is nomologically necessary So, μ cannot be predicate-reduced by any of the π_i But how does it follow, as is required to support the thesis of nonreducibility, that mentalistic predicate μ cannot be reduced by any physicalistic predicate? It does not follow In order to derive the thesis of nonreducibility an additional premise is needed, namely, that, if none of π_1, π_2, π_n predicate reduces μ, then there is no *other* physicalistic predicate π^* that predicate reduces μ If this additional, and rather strong, premise is granted, then the thesis of nonreducibility follows

---

[7] See Place, "Is Consciousness a Brain Process?" *British Journal of Psychology*, XLVII (1956) 44-50, and Smart, "Sensations and Brain Processes," *The Philosophical Review*, LXIIX (1959) 141-56

---

[8] In the terminology of Sydney Shoemaker, P_1, P_2, and so on are *total realizations* of M—"Some Varieties of Functionalism," *Philosophical Topics*, xii (1981) 83-118 Physicalism commits one to the existence of such total realizations Shoemaker correctly distinguishes *total realizations* from what are usually taken to be realizations, which he calls *core realizations* Core realizations are physical properties that *when combined with certain structural properties* realize a mental property For example, according to philosophical lore, the physical property *undergoing C-fiber stimulation* when instantiated in a system with the structure of the human central nervous system realizes the mental property *pain* But *undergoing C-fiber stimulation*, when it is instantiated in a *different* sort of physical structure, is not sufficient for pain Thus, *core realizations* of mental properties are, strictly speaking, *not* realizations of mental properties

[9] Suppose, for reductio, that M is identical to P_1 By hypothesis, if an object o instantiates P_2, then it instantiates M But M is identical to P_1 So, contra the assumption of metaphysical incompatibility, o instantiates both P_1 and P_2 So M is not identical to P_1 Note that a more detailed statement of the argument would require a relativization to times, as an object o might instantiate P_1 at one time, yet not at another

Here is where the reductionist can employ the "disjunctive strategy" The reductionist can claim that mentalistic predicate μ is reduced by the *exhaustive disjunctive predicate* $(\pi_1 \vee \pi_2 \vee \ \pi_n)$ where each π_i designates P_i, and all the P_i designated by some disjunct π_i *exhaust* the realization base of M In other words, the reductionist can reject the rather strong additional premise, and maintain that there is some physicalistic predicate that designates M, namely, the disjunctive predicate $(\pi_1 \vee \pi_2 \vee \ \pi_n)$ The reductionist who employs the disjunctive strategy thus maintains that a disjunctive bridge principle of the form

$$(\vee\text{BP})\ \forall x\,[\mu x \leftrightarrow (\pi_1 x \vee \pi_2 x \vee \ \pi_n x)]^{10}$$

is sufficient for the predicate reduction of μx by $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$

II FODOR, KIM, AND THE THREAT OF RAMPANT ILLEGITIMACY

I shall review here the debate between Fodor and Jaegwon Kim[11] concerning the disjunctive strategy I shall argue that neither Fodor's nor Kim's position is acceptable Fodor argues that, despite the necessary truth of bridge principles such as $(\vee\text{BP})$, the reductionist's disjunctive strategy fails This because disjunctive predicates such as $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ fail to designate legitimate "scientific kinds," though mentalistic predicates such as μx do designate legitimate "scientific kinds"[12] Fodor concludes that psychology and the other

---

[10] The modal strength of such disjunctive bridge principles depends upon the modal strength of the realization relation For example, if 'P realizes F' entails 'If object or event o instantiates P, then it is *metaphysically necessary* that o instantiates F', then such disjunctive bridge principles hold with metaphysical necessity Thus, such disjunctive bridge principles will hold with *at least* nomological necessity, as nomological necessity is, I assume, the weakest grade of modality in terms of which the realization relation could be plausibly defined

[11] See Kim, "Multiple Realization and the Metaphysics of Reduction," in *Supervenience and Mind* (New York Cambridge, 1993), pp 309-35, and Fodor

[12] A predicate might fail to designate a legitimate property, or in Fodor's terms, fail to determine a scientific kind, either because it fails to designate a property at all, or because it designates a property that is in some way illegitimate or unscientific It is not clear to me what a real, yet illegitimate, property would be, and thus I do not see what the motivation for allowing for real, yet illegitimate, properties could be It seems to me that the theoretical role played by a real, yet illegitimate, property would be better played by a *concept* that does not correspond to a real property Kim seems to agree with me on this point, as he seems not to allow for real, yet illegitimate properties (*op cit*, pp 334-35) But, unfortunately, Fodor (*op cit*) seems to allow for real yet illegitimate properties, properties that do not correspond with "kinds" At any rate, the distinction between existent yet illegitimate properties, and nonexistent properties is not central to my concerns, and thus I shall ignore it as much as is possible

"special sciences" are autonomous in the sense that the laws of these special sciences cannot be reduced to the laws of physics. Kim responds by arguing that, given the necessary truth of ($\vee$BP), Fodor cannot coherently accept μx as designating a legitimate kind *and* reject $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ on the grounds that it does *not* designate a legitimate kind; as Kim puts it: "It is difficult to see how one could have it both ways—that is, to castigate $[(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)]$ as unacceptably disjunctive while insisting on the integrity of $[\mu x]$ as [designating] a scientific kind" (*op. cit.*, p. 324). But Kim agrees with Fodor that $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ does not designate a legitimate kind, and thus Kim concludes that μx *also* fails to designate a legitimate kind. I shall argue that Kim is right to criticize Fodor; Fodor's reasons for claiming that $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ is unacceptable while μx is acceptable are confused, and thus Kim is right to insist that the two predicates stand or fall together. But I shall argue that the price of agreeing with Kim that the two predicates fall together is extremely high; it is the price of denying the legitimacy of all, or almost all, properties. Therefore, neither Fodor's nor Kim's position is acceptable.

Fodor objects to the disjunctive strategy on the grounds that disjunctive predicates such as $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ do not designate "scientific kinds." Suppose that there is a (perhaps nonstrict) law of psychology of the form:

(PL) $\Box\forall x\,(\mu_1 x \rightarrow \mu_2 x)$

where '$\Box$' designates a nomological necessity operator, and $\mu_1 x$ and $\mu_2 x$ are atomic mentalistic predicates. (Perhaps (PL) is a law in the "ideal" theory of psychology.) And now consider the disjunctive bridge principles which the reductionist utilizing the disjunctive strategy claims would serve to predicate reduce $\mu_1 x$ and $\mu_2 x$:

(BP1) $\Box\forall x\,[\mu_1 x \leftrightarrow (\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)]$
(BP2) $\Box\forall x\,[\mu_2 x \leftrightarrow (\pi'_1 x \vee \pi'_2 x \vee \ldots \pi'_n x)]$

Given (BP1) and (BP2), the following *physicalistic* law is equivalent to (PL):

(PL*) $\Box\forall x\,[(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x) \rightarrow (\pi'_1 x \vee \pi'_2 x \vee \ldots \pi'_n x)]$

If (PL) is a law (and Fodor thinks that there are, or at least could be, such laws), then it would seem that (PL*) is also a law. (There is no problem substituting inside the nomological necessity operator '$\Box$', as (BP1) and (BP2) guarantee that $\mu_1 x$ and $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$, and $\mu_2 x$ and $(\pi'_1 x \vee \pi'_2 x \vee \ldots \pi'_n x)$, are *at least* nomologically coextensive.) But if (PL*) is a law, then the reductionist carries the day, for from physicalistic law (PL*) together with bridge principles (BP1) and (BP2), psychological law (PL) can be derived, and in this way psychology could be reduced to a more basic physicalistic theory

How does Fodor propose to preclude the purported (law) reduction of (PL) to (PL*)? Fodor claims that the predicates constituting the antecedent and consequent of a law must be "kind predicates " A "kind predicate" for Fodor is, apparently, a *previously formulated* predicate that appears in the previously formulated laws of some true theory Thus, as Fodor would have it, $\mu_1 x$ and $\mu_2 x$ are "kind predicates" as, by assumption, they appear in previously formulated laws of a true psychological theory And consequently (PL) is a law, as the predicates constituting its antecedent and consequent are "kind predicates " But, as $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ and $(\pi'_1 x \vee \pi'_2 x \vee \ \pi'_n x)$ are not previously formulated predicates that appear in a previously formulated law of some true theory, they are not "kind predicates" and, as a result, (PL*) is not a law So psychology cannot be reduced to a physicalistic theory by deriving laws such as (PL) from disjunctive physicalistic "generalizations" such as (PL*) [13]

The upshot is that Fodor requires for the reduction of mentalistic predicate μx a previously formulated physicalistic predicate that is (at least) nomologically coextensive with μx In other words, Fodor is more-or-less requiring for the reduction of psychology (biology, and the like) to physics that the theories of these special sciences turn out to be notational variants of part of the reducing physicalistic theory The chances of this sort of correlation obtaining between theories in the special sciences and physics are, as Fodor suggests, slim But this ought not console the defender of NRP, for Fodor's requirements for reduction are extremely strong, and go far beyond anything required

---

[13] Fodor also objects to the "lawhood" of (PL*) on the grounds that "'it's a law that ____' defines a non truth functional context" (*op cit* , p 140) Fodor argues as follows " one may not argue from 'it's a law that P brings about R' and 'it's a law that Q brings about S' to 'it's a law that P or Q brings about R or S' [F]or example, it is a law that the irradiation of green plants by sunlight causes carbohydrate synthesis, and it is a law that friction causes heat, but [it is not] a law that (either the irradiation of green plants by sunlight or friction) causes (either carbohydrate synthesis or heat)" (*ibid*, p 140)

Fodor is correct that "'it's a law that ____' defines a non truth functional context," and thus the *general* inference Fodor describes is invalid But this does not block the disjunctive strategy In order to block the disjunctive strategy, Fodor must do more than show that the inference from 'it's a law that P brings about R' and 'it's a law that Q brings about S' to 'it's a law that P or Q brings about R or S' is *in general* invalid To block the disjunctive strategy, he must show that this inference is invalid when P and Q exhaust the realization base of a property M_1, and R and S exhaust the realization base of M_2, and it is a proper law that M_1 brings about M_2 Fodor has given us no reason to be suspicious of the inference in this very special circumstance by Nagel's classic model of reduction, or any other model of reduction I know of [14] Moreover, there are several independent reasons for rejecting Fodor's requirement that for μx to be reduced, it must be reduced by a previously formulated physicalistic predicate

First, as was explained above, the issue of predicate reduction is—at least in the context of the mind-body problem—really the issue of property identity If reduction is to have the ontological consequences that the type-type identity theorist desires and the defender of NRP fears, then reducing bridge principles must entail, or at least warrant the assertion of, property identities Thus, what matters for predicate reduction is what the predicates designate, *not* that they have been previously formulated So Fodor cannot reject the disjunctive predicate $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ merely on the grounds that it has not previously been used to state scientific laws The *familiarity* of the predicate is irrelevant, what is relevant is what property, if any, the predicate designates

At this point, one might object that my interpretation of Fodor's reason for rejecting $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ is uncharitable For perhaps Fodor rejects such disjunctive predicates not merely because they have not previously been used to state scientific laws, but rather because such predicates fail to designate legitimate properties or "natural kinds " This objection has some merit, in some passages, it seems that this is what Fodor has in mind [15] The problem with this interpretation of Fodor, however, is that it plays directly into the hands of Kim's response

---

[14] Kim also criticizes Fodor on the grounds that he requires too much for reduction (see *op cit*, footnote 21) In footnote 2 of "Special Sciences," Fodor claims that he is working with what he takes to be the "classical form" of reduction, though he also admits that it is "a stronger one than many philosophers of science hold " He also claims that his argument against reductionism would hold against even "liberalized versions" of reduction I take myself to have demonstrated that this latter claim is false

[15] Fodor introduces the phrase 'kind predicate' with regard to "(natural) kinds" (*op cit*, p 132), thus suggesting that a predicate is an acceptable "kind predicate" if and only if it designates, or "determines," a "natural kind," where the criteria for being a natural kind are independent of any linguistic entity such as a science and/or theory Moreover, he seems to think that disjunctive predicates do not designate "natural kinds", at one point Fodor states, "I doubt that 'is either carbohydrate synthesis or heat' is plausibly taken to be a kind predicate" (*op cit*, p 140) But Fodor later drops the phrase 'natural kind' and rejects the reductionist's appeal to the disjunctive predicate $(\pi_1 x \vee \pi_2 x \ \ \pi_n x)$ on the grounds that it "is not a kind predicate *in the reducing science*" (*op cit*, p 138) In other words, Fodor seems to begin with the plausible requirement for reduction that the predicates in a bridge principle must designate "natural kinds," but then slides to the implausibly strong requirement that the predicates in a bridge principle be previously formulated predicates used to state the laws of the reducing science and/or theory

> If pain is nomically equivalent to *N*, the property claimed to be wildly disjunctive and obviously nonnomic, *why isn't pain itself equally heterogeneous and nonnomic as a kind*? Why isn't pain's relationship to its realization bases, *Nb*, *Nr*, and *Nm*, analogous to jade's relationship to jadeite and nephrite? If jade turns out to be nonnomic on account of its dual "realizations" in distinct microstructures, why doesn't the same fate befall pain? After all, the group of actual and nomologically possible realizations of pain, as they are described by the MR enthusiasts with such imagination, is far more motley than the two chemical kinds comprising jade (*op cit*, p 323)

Kim's criticism then, is this if Fodor is interpreted as advancing the plausible requirement for reduction that the predicates in a bridge principle designate legitimate properties, or "natural kinds," then he cannot reasonably assert that mentalistic predicate μx designates a natural kind, while denying that the exhaustive disjunctive *predicate* $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ designates a natural kind For, given the necessary truth of

$$(\vee\text{BP})\ \forall x\,[\mu x \leftrightarrow (\pi_1 x \vee \pi_2 x \vee \pi_n x)]$$

if $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ fails to designate, or determine, a natural kind, then so does μx Moreover, Kim argues that there is good reason to deny that heterogeneous disjunctive predicates such as $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ designated natural kinds Kim does not object to such predicates merely because they are unfamiliar, rather, he objects to both μx and $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ for semantic and metaphysical reasons "Given that mental kinds are realized by diverse physical causal kinds, it follows that mental kinds are not causal kinds, and hence are disqualified as proper scientific kinds Each mental kind is sundered into as many kinds as there are physical realization bases for it, and psychology as a science with disciplinary unity turns out to be an impossible project" (*op cit*, p 327)

In the next section, I defend the legitimacy of "disjunctive properties" from arguments such as Kim's Here, I wish only to motivate my defense by highlighting an unsavory consequence of Kim's argument It is clear that his argument against the legitimacy of mental properties generalizes into an argument for the illegitimacy of all multiply realized properties, it is irrelevant to the argument that *M* is a *mental* property—all that matters is that *M* is multiply realized For every multiply realized property designated by some predicate ρx there is, or there could be, a corresponding exhaustive disjunctive predicate $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ that is *at least* nomologically coextensive with ρx So, if no disjunctive predicate of the form $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ determines a "scientific kind" or designates a legitimate property, then no predicate such as *px* designates a legitimate property But this is very problematic because all, or almost all, properties are multiply realized If the fact that many different sorts of creatures can instantiate *being in pain* leads us to believe that *being in pain* is multiply realized, then the fact that many different sorts of things can instantiate, for example, *being green* ought to lead us to believe that *being green* is multiply realized The point also holds for paradigmatic physical properties Consider the paradigmatic physical property of having a mass of two grams This property is instantiated by many different kinds of objects—bits of paper, bone, metal, jelly, and so on Moreover, we expect that, if an object (or event) *o* instantiates a property *P*, then there is some noncausal explanation as to why *o* instantiates *P*, there must be something about *o* in virtue of which *o* instantiates *P* That is, there must be some property *R* of *o* such that *R* realizes *P* (So much is required by the "principle of sufficient reason," at least under one formulation of that principle) For example, if a bit of plastic has a mass of two grams, then it instantiates that property in virtue of being composed of a certain number of certain sorts of molecules structured in a certain way And if some very different sort of object *o′*, say a bit of bone, also has a mass of two grams, then *o′* instantiates this property in virtue of being composed of a certain number of a certain other sorts of molecules structured in some other way Claiming that a property *P* is *not* multiply realized is claiming that there are no properties in virtue of which objects instantiate *P*, it is tantamount to claiming that when an object instantiates *P* it is just a "brute fact" that it does so But most philosophers are resistant to positing brute, unexplainable, facts, and to the extent that they are resistant to brute facts, they ought to be accepting of the claim that most if not all properties are multiply realized[16]

---

[16] One might object to the theses that *most* properties are multiply realized on the grounds that *only* "higher-order" properties are multiply realized This objection, however, fails to preclude rampant illegitimacy because, as the notion of a "higher-order property" is usually defined, almost all properties are higher-order properties Louise Antony and Joseph Levine define a *higher-order property* as "a property you have in virtue of having some other property that meets certain specifications"—see "Reduction with Autonomy," *Philosophical Perspectives*, XI (1997) 83-105, here p 85 The problem is that almost every property is associated with some sort of causal/functional role, and thus meets this criterion Similarly, one can apply the Ramsey/Lewis method to define almost any predicate For similar accounts of functional/higher-order properties, see Kim "The Mind-Body Problem Taking Stock After Forty Years," *Philosophical Perspectives*, XI (1997) 185-205, Block "Can the Mind Change the World?" in George Boolos, ed , *Meaning and Method Essays in Honor of Hilary Putnam* (New York Cambridge, 1990), pp 137-70, and Putnam

If the above considerations are correct and thus most, if not all, properties are multiply realized, then the advocate of NRP has a serious problem For if, following Fodor, she rejects the reductionist's appeal to exhaustive disjunctive predicates on the grounds that such predicates do not designate legitimate properties, then she is evidently required to deny the legitimacy of all multiply realized properties The price of rejecting the reductionist's appeal to "disjunctive properties" is rampant illegitimacy The entire argument is explicitly formulated as follows

(1) For each multiply realized property R designated by predicate ρx, there is an exhaustive disjunctive predicate $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ such that each $\pi_i x$ designates P_i where P_i realizes R, and $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ contains one disjunct for each property that realizes R[17]

(2) ρx is *at least* nomologically coextensive with $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$

But since

(3) If ρx and $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ are (at least) nomologically coextensive, then ρx designates a legitimate property only if $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ designates a legitimate property

it follows that

(4) ρx designates a legitimate property only if $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ designates a legitimate property

But in order to undermine the disjunctive strategy and thereby avoid reductionism, the defender of NRP has claimed, following Putnam and Fodor, that

(5) Such heterogeneous disjunctive predicates as $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ do not designate legitimate properties

Therefore

(6) ρx does not designate a legitimate property

But

(7) All, or almost all, properties are multiply realized

And consequently

(8) Few, if any, of our predicates designate legitimate properties

---

[17] This premise presupposes that all the properties in R's realization base can be designated by predicates This assumption probably must be rejected by anyone who is a realist about properties, and this suggests an alternative response to the argument Related issues will be discussed in section IV

The conclusion, (8), is unacceptable, and the argument is valid Moreover, I have argued that the only dubious premise is (5) Thus there is cogent reason for hoping that at least some exhaustive disjunctive predicates such as $(\pi_1 x \vee \pi_2 x \vee \pi_n x)$ do designate legitimate properties (or at least they would if they were formulated), the price of denying this is that of accepting the unsavory conclusion that all, or almost all, properties are illegitimate If rampant illegitimacy is to be avoided, an acceptable account of disjunctive properties must be formulated

III DEVELOPMENT OF A CAUSAL POWER ACCOUNT OF DISJUNCTIVE PROPERTIES

Before an acceptable account of disjunctive properties can be formulated, it must be made clear what it would be for a property to be a *disjunctive* property I have been, and shall continue to be, assuming a naive realism concerning properties, properties are mind-independent entities, at least some of which are designated by predicates A disjunctive property, one quickly thinks, is what one gets when one takes property *P* and property *Q* and then forms their disjunction But this account involves a use/mention confusion one can form the disjunction of predicate ϕ and predicate ψ, namely, $\ulcorner \phi \vee \psi \urcorner$, but one cannot literally form a disjunction of properties any more than one could form a disjunction of, say, particular people [18] Let us define the notion of a *properly disjunctive predicate* a disjunctive predicate $(\pi_1 \vee \pi_2 \vee \pi_n)$ is a properly disjunctive predicate if and only if (i) there is more than one disjunct π_i, (ii) each disjunct π_i designates a legitimate property, and (iii) each π_i designates a distinct property The notion of a *disjunctive property* is now defined as follows *P* is a disjunctive property if and only if *P* can be designated by a properly disjunctive predicate (In terms of the lambda calculus, *P* is a disjunc-

---

[18] In a puzzling passage in "Concepts of Supervenience," in *Supervenience and Mind,* Kim eschews infinite disjunctive predicates, yet endorses the "operation" of disjunction applied to infinitely many properties " such operations as infinite conjunctions and infinite disjunctions would be highly questionable for predicates, but not necessarily for properties—any more than infinite unions and intersections are for classes The property of being less than one meter long can be thought of as an infinite disjunction (e g of all properties of the form being less than n/n+1 meters long, for every natural number n)" (p 73)

I share Kim's uneasiness concerning infinite predicates But, while it is relatively clear what a disjunction of predicates is, it is not at all clear what a disjunction of properties is tive property if and only if for some predicates ϕ and φ designating *distinct* legitimate properties, P is designated by $\ulcorner \lambda x(\phi x$ or φx or $) \urcorner$[19]

Thus the nonreductionist's claim that there are no disjunctive properties (premise (5) in the above argument) is interpreted to mean that properly disjunctive predicates do not designate legitimate properties, such predicates are either "empty" or designate illegitimate, "unscientific," properties Note that the exhaustive disjunctive predicates by which the reductionist proposes reducing mentalistic predicates are properly disjunctive predicates, they satisfy conditions (i)-(iii) above Consequently, since all or almost all properties are multiply realized, the nonreductionist's claim that there are no legitimate disjunctive properties entails rampant illegitimacy

My strategy for formulating an acceptable account of disjunctive properties will be to explicate and criticize two of David Armstrong's[20] influential arguments against the legitimacy of disjunctive properties These negative arguments serve as constraints on an acceptable account of disjunctive properties in the sense that—at the cost of rampant illegitimacy—an acceptable account of disjunctive properties must allow at least some disjunctive properties to survive these negative arguments Thus, in responding to these arguments, I shall be taking significant steps toward the development of an acceptable account of disjunctive properties

---

[19] One could also provide a nonmetalinguistic account of *disjunctive properties* as follows let properties be functions from possible worlds to sets of entities in those worlds, that is, property P is a function $f(\ )$ such that for worlds w, $f(w) = \{x\ \ x$ has P in $w\}$ The disjunction $f_d(w)$ of properties $f_1(\ )$ and $f_2(\ )$ is defined as follows for all w, $f_d(w) = f_1(w) \cup f_2(w)$ A property could then be defined as a *disjunctive property* as follows property P is *disjunctive* if and only if there are properties Q and R such that P is the disjunction of Q and R To avoid the result that all properties are disjunctive, various restrictions could be placed on Q and R, for example, it might be required that Q and R be "natural" or "basic" properties

The problem with this account of disjunctive properties, aside from that of specifying the "natural" properties, is that it is not clear how to relate it to the debate surrounding the argument from multiple realizability Note that prima facie there is no reason to suppose that disjunctive predicates such as $(\pi_1 x \vee \pi_2 x \vee\ \ \pi_n x)$ designate disjunctive properties in this sense Nor is there reason to suppose that atomic predicates do not designate such disjunctive properties

[20] *A Theory of Universals Universals and Scientific Realism, Volume II* (New York Cambridge, 1978) For other arguments against disjunctive properties, see D H Mellor, "Predicates and Properties," in Mellor and Alex Oliver, eds , *Properties* (New York Cambridge, 1997), David Owens, "Disjunctive Laws," *Analysis*, IL (1989) 197-202, and Kim, "Multiple Realization and the Metaphysics of Reduction," in *Supervenience and Mind*, pp 309-35 Mellor's argument is successfully criticized by Andrew Botterell in "Mellor on Negative Properties," *The Philosophical Quarterly*, XLVIII, 193 (1998) 523-26 Owens and Kim's argument is successfully criticized by Fodor in "Special Sciences Still Autonomous After All These Years," in *In Critical Condition Polemical Essays on Cognitive Science and the Philosophy of Mind* (Cambridge MIT, 1998), pp 11-12

Armstrong's first argument against disjunctive properties is this

> disjunctive properties offend against the principle that a genuine property is identical in its different particulars Suppose *a* has a property P, but lacks Q while *b* has Q but lacks P It seems laughable to conclude that from these premises that *a* and *b* are identical in some respect Yet both have the "property", P *or* Q (*op cit*, p 20)

This argument seems persuasive when applied to the sort of extremely "heterogeneous" disjunctive predicate Armstrong had in mind Consider his favorite example 'is a raven or a writing desk' It would be at least odd to say that the entities in the extension of 'is a raven or a writing desk' all resemble each other, or "are identical in some respect "[21] Thus the argument shows that not *all* properly disjunctive predicates designate properties But it by no means follows that *no* properly disjunctive predicates designate properties Consider, for example, the (perhaps infinite) disjunctive predicate $(\chi_1 x \vee \chi_2 x \vee \quad \chi_n x)$ where each disjunct $\chi_i x$ designates a distinct color property, C_i, and *every* color is designated by some χ_i (Thus $(\chi_1 x \vee \chi_2 x \vee \quad \chi_n x)$ is something like, 'is either red or blue or green or yellow or orange ') This predicate satisfies requirements (i)-(iii) for being a *properly disjunctive predicate* Thus, if it designates a property, then, by the above definition of disjunctive property, it designates a disjunctive property The property *being blue* (or being a very particular shade of blue) is one of the C_i, and *being blue* is a *determinate* of the *determinable* property *being colored*—*being blue* is one of the ways in which a thing can be colored, and if a thing is blue, then it *must* be colored To put the point in more relevant terminology, *being blue realizes being colored*, more generally, *determinables* with more than one *determinate* are one sort of *multiply realized* property [22] Consequently, because the properly disjunctive predicate $(\chi_1 x \vee \chi_2 x \vee \quad \chi_n x)$ is *exhaustive*, it is necessarily true that it is satisfied by all and only colored objects Therefore, assuming that the entities that instantiate *being colored* all resemble each other in some respect, there is a *respect* in which all objects in the extension of $(\chi_1x \vee \chi_2x \vee \ \chi_nx)$ resemble each other, namely, *being colored*

---

[21] It would not, however, be false to say this For example, all ravens and all writing desks have mass, and thus are identical in this respect But it suits my explanatory purposes to grant this point to Armstrong, at least with regard to his first argument

[22] The isomorphism between realized and realizing properties, on the one hand, and determinables and determinates, on the other, was, I believe, first noticed by Stephen Yablo, who utilizes the isomorphism to explain the causal relevance of mental events—"Mental Causation," *The Philosophical Review*, CI (1992) 245-80 My views concerning disjunctive properties are significantly influenced by Yablo's insights

The same response, mutatis mutandis, illustrates that the exhaustive properly disjunctive predicate $(\pi_1x \vee \pi_2x \vee . \ \pi_nx)$, by which the reductionist proposes reducing mentalistic predicate μx, also serves as a counterexample to Armstrong's first argument For μx and $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ are also necessarily coextensive Consequently, there is a "respect," namely, *M*, in which all the entities in the extension of $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ resemble each another As is required to steer clear of rampant illegitimacy, this most relevant case serves as a counterexample to Armstrong's first argument against disjunctive properties, and thus his argument does not demonstrate that all properly disjunctive predicates fail to designate legitimate properties

Although Armstrong's first argument fails to support the nonreductionist's rejection of disjunctive properties, it does serve to highlight a significant constraint on disjunctive properties His first argument fails to demonstrate that $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ designates a legitimate property because the π_ix which are disjoined to form $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ *overlap* on a property Consider again all specific color predicates and the determinable *being colored* All specific color predicates can be said to overlap on the property *being colored* because an object that satisfies any one of the specific color predicate must instantiate *being colored* Generalizing now, the disjuncts of a disjunctive predicate can be said to *overlap* if and only if there is some property *R* such that every possible object (or event) that satisfies *any* of the disjuncts *must* instantiate *R* It is because its disjuncts overlap on *M* that $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ survives Armstrong's first argument because the disjuncts overlap on *M*, there is a "respect" in which all the objects (or events) in the extension of $(\pi_1x \vee \pi_2x \vee \ \pi_nx)$ resemble each other, namely, *M* Although Armstrong's first argument does not succeed in demonstrating that there are no disjunctive properties, it does succeed in demonstrating that a necessary condition for a properly disjunctive predicate's designating a legitimate property is that the disjuncts of the predicate overlap on a property

Armstrong's second argument against disjunctive properties is this

> the postulation of disjunctive properties breaks the link which it is natural to make between properties of things and causal powers of things Suppose that *a* has P but lacks Q The predicate 'P∨Q' applies to *a* Nevertheless, when *a* acts, it will surely act only in virtue of its being

> P Its *being P or Q* will add no power to its arm This suggests that *being P or Q* is not a property (*op cit*, p 20)[23]

This argument presupposes a causal power conception of properties, the essentials of which Armstrong describes in this passage.

> (a) The active and passive powers of particulars are determined by their properties (b) Every property bestows some active and/or passive power[s] upon the particulars of which it is a property (c) A property bestows the very same causal power[s] upon any particular of which it is a property (d) Each different property bestows a different [set of] power[s] on the particulars of which it is a property (*op cit*, pp 43-44)

(I have amended Armstrong's conception slightly to allow properties to bestow nonempty sets of causal powers rather than just single causal powers It is not clear whether or not Armstrong would accept this amendment[24])

Armstrong does not extend this conception of properties as bestowing causal powers to an account of the conditions under which a predicate designating a legitimate property is satisfied by an object, but it is relatively clear how the conception ought to be so extended Legitimate properties bestow causal powers on the objects that instantiate them, an object *o* instantiates a legitimate property *P* if and only if *o* possesses every causal power bestowed by *P* It will simplify matters if, instead of speaking of properties "bestowing" causal powers, properties are simply identified with sets of causal powers Thus, I shall sometimes speak of a property being *constituted by* a set of causal powers The causal power model of properties together with this simplifying assumption implies the following necessary and jointly sufficient conditions for a predicate's designating a property

> Predicate π designates a property *P if and only if* there is some nonempty set of causal powers *p* such that (a) if a particular *o* satisfies π then *o* possesses every power in *p*, and the converse (b) if a particular *o* possesses every power in *p*, then *o* satisfies π

Armstrong's second argument again is cogent when applied to extremely heterogeneous properly disjunctive predicates such as his example, 'is a raven or a writing desk' The problem is that this predicate does not meet conditions (a) and (b). That is, there is no set of causal powers *d* such that (a) every particular that satisfies 'is a raven or a writing desk' possesses every causal power in *d*, and (b) every particular that possesses every causal power in *d* satisfies 'is a raven or a writing desk'. Let *r* be the set of causal powers associated with the predicate 'is a raven', and *w* be the set of causal powers associated with 'is a writing desk'. Set of causal powers *d* cannot be identified with *r*, because writing desks satisfy 'is a raven or a writing desk' yet do not possess all the causal powers in *r*. Thus condition (a) is violated. Nor can *d* be identified with *w*, for ravens satisfy 'is a raven or a writing desk' yet do not possess all the causal powers in *w*. Thus condition (a) is again violated. And a fortiori *d* cannot be identified with $r \cup w$, because many ravens satisfy 'is either a raven or a writing desk' but do not possess all the powers in $r \cup w$, and many writing desks also satisfy 'is a raven or a writing desk' but do not possess all the powers in $r \cup w$. Clearly, then, in order to satisfy condition (a), *d* must be a subset of *both* *r* and *w*. Perhaps then *d* can be identified with $r \cap w$. Note that it is unlikely that $r \cap w$ is the empty set. For there are *some* causal powers shared by all possible ravens and writing desks. For example, surely every possible raven or writing desk has mass, and *having mass* is a property, and thus everything that instantiates this property must possess every member of a nonempty set of causal powers. So $r \cap w$ is not the empty set. But nonetheless *d* cannot be identified with $r \cap w$. The problem is that it is likely that there are many possible things that possess all the causal powers in $r \cap w$, but do not satisfy the predicate 'is either a raven or a writing desk'. If an object *o* possesses all the causal powers that all ravens and writing desks have in common, it does not follow that *o* is either a raven or *o* is a writing desk. There are, for example, many things that possess all the causal powers bestowed by the property *having mass* that are neither ravens nor writing desks. And thus *d* cannot be identified with $r \cap w$ on pain of condition (b) being violated. Moreover, and a fortiori, *d* cannot be identified with any proper subset of $r \cap w$. If there are objects that possess every causal power in $r \cap w$ yet do not satisfy 'is either a raven or a writing desk', then for any proper subset of $r \cap w$ there will be objects that possess every causal power in it, but do not satisfy 'is either a raven or writing desk'.

---

[23] This is actually Armstrong's third argument against disjunctive properties, but his second argument is neither successful nor relevant to my purposes

[24] These amendments bring Armstrong's causal power model of properties more in line with the model proposed by Shoemaker, see "Causality and Properties," in Peter van Inwagen, ed, *Time and Cause* (Boston Reidel, 1980), pp 229-54, and "Causal and Metaphysical Necessity," *Pacific Philosophical Quarterly*, LXXIX (1998) 59-77 My views on these matters owe much Shoemaker's work

So Armstrong's second argument, like his first argument, succeeds in demonstrating that 'is either a raven or a writing desk' fails to designate a legitimate property. But his second argument, again like his first argument, does not demonstrate that *no* properly disjunctive predicate designates a property. Consider again the exhaustive disjunction of color predicates, $(\chi_1 x \vee \chi_2 x \vee \ldots \chi_n x)$, where each disjunct

$\chi_i x$ designates a distinct color property, C_i, and each such property C_i is constituted by a set of causal powers c_i If $(\chi_1 x \vee \chi_2 x \vee \chi_n x)$ designates a legitimate property, there must be some nonempty set of causal powers d' such that (a) every possible thing that satisfies $(\chi_1 x \vee \chi_2 x \vee \chi_n x)$ possesses every causal power in d', *and* (b) every possible thing that possesses every causal power in d' satisfies $(\chi_1 x \vee \chi_2 x \vee \chi_n x)$ Is there a set of causal powers that meets conditions (a) and (b)? In this case, there is good reason to suppose that there is Specifically, there is good reason to suppose that the intersection of the all the sets of causal powers constituting the properties designated by the disjuncts, namely, $c_1 \cap c_2 \cap c_n$, meets both conditions

First, assuming the causal power model of properties, it is incontestable that $c_1 \cap c_2 \cap c_n$ meets condition (a) Suppose an object o satisfies $(\chi_1 x \vee \chi_2 x \vee \chi_n x)$ Then o must satisfy one of the disjuncts $\chi_i x$ But if o satisfies one of the disjuncts $\chi_i x$, then it must instantiate the property C_i designated by that disjunct, and thus it must possess every causal power in the set of causal powers c_i that constitute C_i But if o has every causal power in some set c_i, then o necessarily has every causal power in $c_1 \cap c_2 \cap c_n$ So $c_1 \cap c_2 \cap c_n$ meets condition (a)

It is not incontestable that $c_1 \cap c_2 \cap c_n$ meets condition (b) because the argument that it does so presupposes a particular account of the *realization* relation This account of realization, however, is strongly suggested by the causal power model of properties According to the working definition of *realization,* a property P *realizes* a property Q if and only if (i) if an object (or event) o instantiates P, then, necessarily, o instantiates Q, and (ii) o's being P explains—in a metaphysical sense—o's being Q, being P is one of the *ways in which* a thing can be Q On the causal power model of properties, an object instantiates a property if and only if it possesses every causal power in the set that constitutes that property So, putting these ideas together leads naturally to the following general definition of *realization*

> *P realizes Q* if and only if (def), where p and q are the sets of powers constituting P and Q, $q \subset p$

Consider the relation that obtains between the determinate property *being yellow* and the determinable of that determinate, *being colored* Because property P's being a determinate of a determinable property Q is sufficient for P's *realizing Q, being yellow* realizes *being colored* On the causal power model of properties, the properties *being yellow* and *being colored* are constituted by sets of causal powers Let us call these sets of causal powers y and c, respectively Now consider all the things that instantiate *being yellow* These things are yellow in virtue of having *all* the causal powers in set y And consider all the things that instantiate *being colored* These things are colored in virtue of possessing *all* the causal powers in set c What relation obtains between c and y? Clearly, c is a proper subset of y every causal power something possesses in virtue of instantiating *being colored* is also a power it possesses in virtue of instantiating *being yellow* But not every power a thing possesses in virtue of instantiating *being yellow* is a power it possesses in virtue of instantiating *being colored* (For example, yellow warning signs are visible to drivers at night, but blue, green, and purple signs are not) And consider the relation that obtains between the set of causal powers constituting the mental property M and one of M's realizors, P_1 Let us identify these properties with sets of causal powers m and p_1 Again, consider all the things that instantiate M These things are M in virtue of possessing all the causal powers in m And consider all the things that instantiate some realizing property P_1 These things instantiate P_1 in virtue of possessing all the causal powers in p_1 What relation obtains between m and p_1? Again, it is clear that m is a proper subset of p_1 every power something possesses in virtue of instantiating M is also a power something would posses in virtue of instantiating P_1 But not every power a thing possesses in virtue of instantiating P_1 is a power it would possess in virtue of instantiating M (For example, suppose M is *believing that snow is white* And suppose, as functionalism would have us believe, that this mental property can be instantiated by a machine made of steel, and that only such steel machines instantiate P_1 Such a machine would possess the power of being magnetic in virtue of instantiating P_1, but it would not posses the power of being magnetic in virtue of instantiating *believing that snow is white*)

If one grants the above account of realization, then $c_1 \cap c_2 \quad \cap c_n$ meets condition (b), that is, if an object o possesses every causal power in $c_1 \cap c_2 \quad \cap c_n$, then o satisfies $(\chi_1 x \vee \chi_2 x \vee \quad \chi_n x)$ According to the above account of realization, a multiply realized property is constituted by the intersection of the sets constituting its realizors [25]

---

[25] Let C be the set of causal powers constituting *being colored,* and let c_1 and c_2 and c_n be sets of causal powers constituting specific color properties Suppose C has as a member some causal power *cp* that is not a member of $c_1 \cap c_2 \quad \cap c_n$ Then there would have to be a c_i that did not have *cp* as a member But then something could possess all the causal powers in this c_i yet not possess all the powers in C So, a thing could instantiate the color property constituted by this c_i yet not instantiate *being colored* But this is absurd, as every thing that is some specific color or other is colored So C cannot contain some causal power that is not a member of $c_1 \cap c_2 \quad \cap c_n$ Conversely, suppose $c_1 \cap c_2 \quad \cap c_n$ has as a member some causal power that is not a member of C Then a thing could have every causal power constituting *being colored,* yet lack a causal power that is possessed by every object that instantiates some specific color property or other So, an object could instantiate being colored,

Thus $c_1 \cap c_2 \ldots \cap c_n$ constitutes *being colored*. Therefore, any object *o* that possesses all of the causal powers in $c_1 \cap c_2 \ldots \cap c_n$ instantiates *being colored*. But if *o* instantiates *being colored*, then *o* must instantiate some specific color property P_i. That is, if *o* has all of the causal powers in $c_1 \cap c_2 \ldots \cap c_n$, then *o* must either have all the causal powers in c_1, or have all the causal powers in c_2, or . . . But if *o* possesses all the causal powers in some set c_i, then *o* must satisfy the corresponding disjunct $\chi_i x$, and hence *o* must satisfy $(\chi_1 x \vee \chi_2 x \vee \ldots \chi_n x)$. So, assuming the account of realization that is strongly suggested by the causal power model of properties, $c_1 \cap c_2 \ldots \cap c_n$ also meets condition (b).

The same arguments apply, mutatis mutandis, to the most relevant case of the reductionist's properly disjunctive predicate, $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$. For each realizing physical property P_i designated by a disjunct $\pi_i x$, let p_i be the set of causal powers that constitutes P_i. For the reasons given above, the set $p_1 \cap p_2 \ldots \cap p_n$ meets both conditions. Condition (a) is met because every possible thing that satisfies $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ possesses every causal power in $p_1 \cap p_2 \ldots \cap p_n$, and condition (b) is met because every possible thing that possesses every causal power in $p_1 \cap p_2 \ldots \cap p_n$ satisfies $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$. And consequently, as is required to avoid rampant illegitimacy, it cannot be concluded from Armstrong's second argument that the reductionist's $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$ fails to designate a legitimate property.

Armstrong's second argument, like his first argument, fails to demonstrate that no properly disjunctive predicates designate legitimate properties. But the argument serves to place an additional constraint on an acceptable account of disjunctive properties. A predicate π designates a legitimate property *P* if and only if there is some nonempty set of causal powers *p* such that (a) if a particular *o* satisfies π, then *o* possesses every power in *p*, and the converse, (b) if a particular *o* possesses every power in *p*, then *o* satisfies π. What must be true of a properly disjunctive predicate if it is meet conditions (a) and (b)? The constraint derived from Armstrong's first argument is that the disjuncts of a properly disjunctive predicate must *overlap* on a property. Let us now redefine the notion disjuncts overlapping in terms of the causal power model of properties: the disjuncts of a disjunctive predicate *overlap* on set of powers if and only if every object (or event) that satisfies any of the disjuncts must possess every

---

yet not instantiate some specific color property. But this also is absurd: every thing that is colored is some specific color. Therefore, the set of causal powers constituting *being colored* must be the intersection of all the sets of causal powers constituting specific color properties. And, generalizing now, a multiply realized property is constituted by the intersection of the sets constituting its realizors.

causal power in that set It can now be discerned that it is necessary and sufficient for a properly disjunctive predicate's meeting condition (a) that its disjuncts overlap on a set of causal powers What Armstrong's second argument illustrates is that condition (b) must also be satisfied That is, it is not enough for $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ to designate a legitimate property that its disjuncts overlap on property M, it also must be the case that any object that instantiates M must satisfy $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ Let us say that a properly disjunctive predicate that has both of these features is such that its disjuncts *satisfactorily overlap*

I conclude that Armstrong's arguments demonstrate that properly disjunctive predicates whose disjuncts do not satisfactorily overlap fail to designate legitimate properties Assuming the causal power model of properties and the attendant account of realization, however, a properly disjunctive predicate whose disjuncts satisfactorily overlap does designate a legitimate property These results entail the following general account of disjunctive properties a properly disjunctive predicate designates a legitimate property if and only if its disjuncts *satisfactorily overlap* Moreover, in the most relevant case of the reductionist's exhaustive properly disjunctive $(\pi_1 x \vee \pi_2 x \vee \ \pi_n x)$ where each disjunct $\pi_i x$ designates one of M's realizors, the disjuncts satisfactorily overlap on the intersection of the sets of causal powers constituting M's realizors, namely, $p_1 \cap p_2 \ \cap p_n$ These results are desirable, as they provide a means for avoiding rampant illegitimacy More specifically, because the disjuncts of the reductionist's exhaustive properly disjunctive predicates satisfactorily overlap, premise (5) in the argument presented in section II can be rejected

IV THE CONSEQUENCES FOR NRP

The causal power model of properties and the attendant accounts of disjunctive properties and the realization relation have significant consequences for NRP Recall that NRP is the conjunction of two theses

Physicalism all particulars are constituted by physical particulars, and all properties are realized by physical properties

Nonreducibility mental predicates *cannot* be reduced by physical predicates

The above definition of realization serves to clarify and support the thesis of physicalism First, the definition immediately secures the result, central to NRP, that mental properties are not identical to the physical properties that realize them, for no set is identical to one of its proper subsets Second, the definition helps to clarify that NRP is incompatible with the metaphorical claim, often associated with NRP, that mental properties exist "over and above" their realizors According to the above well-motivated definition, multiply realized mental properties, though real and causally efficacious, are better thought of as *parts* of their physical realizors The definition thus undermines the standard "levels" picture, according to which mental properties exist "at a higher level" than their physical realizors, and consequently it falsifies Kim's equation of NRP with some form of emergentism [26] The falsification of this equation is an extremely beneficial result for NRP, for without this equation the problems forcefully presented by Kim concerning causal and explanatory exclusion of mental properties by physical properties do not arise [27] Just as there is no causal and/or explanatory competition between a whole and its parts, so there is no causal and/or explanatory competition between instances of mental properties and instances of their physical realizors And, finally, the causal power model of properties combined with the above definition of the realization relation surpass the old notions of supervenience in providing an explanation of the relation between mental properties and physical properties Kim[28] has recently complained that "mind-body supervenience is not an explanatory theory, it merely states a pattern of property covariation between the mental and the physical, and points to the existence of a dependency relation between the two Yet it is wholly silent on the nature of the dependency relation that might explain why the mental supervenes on the physical" (*ibid*, p 190) The above definition of realization provides at least the foundation for an explanation of this dependency relation mental properties depend upon physical properties because mental properties in part constitute certain physical properties Of course, all of these remarks are highly programatic, and carrying out the relevant programs would require a more detailed working out of the causal power model of properties than I have attempted here But these remarks do illustrate that advocates of NRP have much to gain from adopting the causal power model of properties and the definition of realization that is implied by that model

Although the causal power model of properties and the attendant account of realization provide some support for and clarification of the thesis of physicalism, they seem to spell trouble for the thesis of nonreducibility. The most prevalent argument in support of nonreducibility is the argument from multiple realizability, which argues that mentalistic predicates such as μ cannot be reduced by any physicalistic predicate because mental properties are multiply realized by physical properties. Against this argument the reductionist can employ the "disjunctive strategy": he can argue that, despite the phenomenon of multiple realization, mentalistic predicates such as μ are reduced to physicalistic predicates by way of disjunctive bridge principles of the form

---

[26] See Kim, "The Nonreductivist Troubles with Mental Causation," in John Heil and Alfred Mele, eds , *Mental Causation* (New York Oxford, 1993), pp 189-210

[27] See Kim (*op cit*), and "The Myth of Nonreductive Materialism," *Proceedings and Addresses of the American Philosophical Association*, LXIII (1989) 31-47

[28] "The Mind-Body Problem Taking Stock After Forty Years," *Philosophical Perspectives*, XI (1997) 185-205

$$(\vee\mathrm{BP})\ \forall x\ [\mu x \leftrightarrow (\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)]$$

The standard nonreductionist reply to the reductionist's "disjunctive strategy" is to "question the propriety of $[(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)]$ as a legitimate property."[29] But the considerations of the previous sections undermine this standard nonreductionist reply, for they provide the reductionist with a well-motivated and plausible account of "disjunctive properties," and on this account the reductionist's exhaustive disjunctive predicates would designate legitimate properties. Thus, unless another reply to the disjunctive strategy can be formulated, rampant illegitimacy is avoided only at the cost of the argument from multiple realizability in support of nonreducibility.

Fortunately, another reply to the disjunctive strategy can be formulated, and the formulation of this reply serves to clarify NRP further. Consider again the bridge principle schema ($\vee$BP), and in particular the disjunctive predicate schema $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$. What exactly is this disjunctive predicate schema schematic of? As is often pointed out, it is likely that mental property *M*, whatever it is, will have an infinite realization base. This will mean that to predicate reduce μ the reductionist must invoke an exhaustive disjunctive predicate that has infinitely many disjuncts. But nobody ever has, or ever will, formulate an infinite predicate. Moreover, nobody has ever even formulated a single predicate $\pi_1 x$ such that for some mental property *M*, the property designated by $\pi_1 x$ realizes *M*. (One reason for this is that the physical realizors of many mental properties are *wide*—they include complex environmental features external to the creature instantiating the mental state.) In short, the reductionist's appeal to exhaustive disjunctive predicates is an idealization of a rather extreme sort. We are familiar with many mentalistic predicates that, it is assumed here, designate mental properties. But we have no idea what the relevant disjunctive predicates might be, and we have good reason to believe that we cannot really formulate such disjunctive predicates. The reductionist's appeal to disjunctive predicates must then be understood as a claim concerning what is *in principle possible*, it is a claim to the effect that some sort of epistemologically ideal being could reduce mentalistic predicates by exhaustive infinite disjunctions of physicalistic predicates. The reductionist is asserting the counterfactual claim that *if we could* formulate infinite physicalistic predicates of the form $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$, then we could reduce mentalistic predicates. Moreover, the results of the previous sections suggest that the reductionist's counterfactual claim is, despite the arguments against disjunctive properties, true. But is the truth of the reductionist's counterfactual claim incompatible with the thesis of nonreducibility?

---

[29] Kim, *The Philosophy of Mind*, (New York: Westview, 1996), p. 219.

It all depends on how the thesis of nonreducibility is interpreted. If the 'cannot' in this thesis is interpreted to be very strong, so that the thesis states that it is not even in principle possible to reduce mentalistic predicates, then the reductionist carries the day. For his counterfactual claim is true: *if* we *could* formulate exhaustive infinite physicalistic predicates of the form $(\pi_1 x \vee \pi_2 x \vee \ldots \pi_n x)$, then we could reduce mentalistic predicates. But if the 'cannot' is interpreted to be less strong, so that the reductionist's counterfactual claim is compatible with nonreducibility, then the defender of NRP carries the day. And this is the appropriate response for the defender of NRP to make against the disjunctive strategy: the defender of NRP should grant that it is in principle possible for mentalistic predicates to be reduced by exhaustive disjunctive predicates, but he should deny that this is incompatible with nonreducibility. This weaker version of NRP does not deny that it is in principle possible for "ideal" scientists to formulate physicalistic predicates that would reduce our mentalistic predicates. (It is difficult to see how such a denial could be compatible with physicalism.) It rather claims that we really shall not and cannot reduce our mentalistic predicates to physicalistic predicates. One might object that by weakening the thesis of nonreducibility in this way, NRP is rendered a purely epistemological and thus uninteresting doctrine. For on this weaker view, the nonreducibility of mentalistic predicates is not explained by mental properties and physical properties being in different ontological categories; there is no reason to assume that mental properties are in any ontologically significant sense on a "higher level," nor that they in any ontologically significant sense "emerge" from physical properties. Rather, the weaker view suggests that the nonreducibility of mentalistic predicates is purely due to our own epistemological limitations. But, first, I do not think this is an objection so much as it is a clarification of what a plausible version of NRP must claim, NRP, properly understood, is primarily an epistemological doctrine. It alleges that the theories and predicates of the special sciences never will in fact be reduced to physical theory. And, second, even this weaker version of NRP ought to satisfy psychologists, economists, and nonreduction-minded philosophers on this weaker conception of NRP, generalizations expressed in terms of mentalistic predicates can be supported by evidence, and can support counterfactuals. Thus there is no special reason to doubt the existence of psychological, economic, and such laws. Consequently, even on this weaker conception, psychology and the other special sciences are guaranteed autonomy from physics, though, like physics, they are by no means guaranteed to be successful.

LENNY CLAPP

Illinois Wesleyan University

THE END OF MILLIANISM MULTIPLE BEARERS, IMPROPER NAMES, AND COMPOSITIONAL MEANING*

Millianism, in J S Mill's[1] words, claims that

> Proper names are not connotative They denote the individuals who are called by them, but they do not indicate or imply any attributes as belonging to those individuals (*ibid*, I, 2, 5, 33)

Mill argued that the proper name 'Dartmouth' cannot be taken to have the connotation (sense) *city at the mouth of the river Dart* because

> If sand should choke up the mouth of the river [Dart], or an earthquake change its course, and remove it to a distance from the town, the name of the town would not necessarily be changed That fact, therefore, can form no part of the signification of the word, for otherwise, when the fact confessedly ceased to be true, no one would any longer think of applying the name (*ibid*, I, 2, 5, 34)

Even though this argument is clear, cogent, and appeared in a widely read book by a highly influential philosopher, it took over a century for the argument to have a significant impact on the philosophy of language

The reason for the delay in such impact is that the philosophers whose work shaped the discussion of names simply took no notice of Mill's argument Gottlob Frege,[2] who paid close attention to everything Mill said about number in Books II and III of *A System of Logic*, ignored everything he said about names in Book I Bertrand Russell[3] developed his theory that proper nouns like 'Socrates' are descriptions without ever mentioning Mill's argument against such theories Ludwig Wittgenstein,[4] who might well have adopted Mill's view, adopts an apparently incompatible one, roughly, that a name is

---

* I wish to thank Bradley Armour-Garb, Kent Bach, Edward Becker, Charles Chihara, Arthur Collins, Ned Block, David Pitt, and Virginia Valian Thanks also to my audiences at Centre de Recherche en Epistémologie Appliquée, the December 1999 New York Conference on Methods, the Simon Fraser University Philosophy Department, and the University of California/Berkeley Philosophy Department

[1] *A System of Logic*, Volume I (London Parker, Son, and Bourn, 1862)

[2] *The Foundations of Arithmetic*, J L Austin, trans (Oxford Blackwell, 1953), pp 29-32

[3] "The Philosophy of Logical Atomism," *The Monist*, xxviii (1918) 523-25

[4] *Philosophical Investigations* (Oxford Blackwell, 1953), § 79

0022-362X/01/9803/137-66

associated with a family of descriptions from which a speaker draws as the occasion demands. Furthermore, we look in vain in *Meaning and Necessity: A Study in Semantics and Modal Logic*, Rudolf Carnap's[5] celebrated reconstruction of Fregean semantics, for what he has to say about Mill's argument. Finally, taking his cue from Wittgenstein's discussion of names, John Searle[6] proposed his "sufficient but so far unspecified number of [descriptions]" view, but, again, there is no mention of Mill's argument. It is not until Saul Kripke's[7] imaginative elaboration of Millian modal reasoning in *Naming and Necessity* that the argument comes into its own. Over the next decades, Kripke's barrage of Mill-style counterexamples succeeded in persuading most philosophers of language of the truth of Millianism.

I. THE PURE METALINGUISTIC THEORY

Over the years, I have been arguing that Millian modal reasoning does not establish the truth of Millianism.[8] Such reasoning shows that classical, Fregean, descriptivism is false, but it does not show that Millianism is true because it does not show that descriptivism is false. The reasoning shows that proper names do not have a Fregean sense, but not that they do not have a sense. The hitch is that the Fregean notion of sense is not the only notion of sense, and, hence, classical descriptivism is not the only form of descriptivism. In addition to classical descriptivism, there is the option of a revisionist, non-Fregean, form of descriptivism.

This option has been neglected every bit as much as Mill's argument, though, to be sure, for a totally different reason. Non-Fregean descriptivism was neglected not because philosophers chose to ignore it, but because, equating sense with Fregean sense, and, hence, intensionalism with Fregean intensionalism, they simply had no inkling of its existence. Thus, the first step in correcting the situation is to provide a non-Fregean definition of sense. This is (D).

---

[5] Chicago: University Press, 1956, Enlarged Edition, Section 79.

[6] "Proper Names," *Mind*, LXVII (1958): 166-73.

[7] "Naming and Necessity," in Donald Davidson and Gilbert Harman, eds., *Semantics of Natural Language* (Boston: Reidel, 1972), pp. 253-355; *Naming and Necessity* (Cambridge: Harvard, 1980). References in this paper are to the latter publication.

[8] "A Proper Theory of Names," *Philosophical Studies*, XXXI, 1 (1977): 1-80; "The Neoclassical Theory of Reference," in Peter A. French, Thomas E. Uehling, Jr., and Howard K. Wettstein, eds., *Contemporary Perspectives in the Philosophy of Language* (Minneapolis: Minnesota UP, 1979), pp. 103-24; "Has the Description Theory of Names Been Refuted?" in George Boolos, ed., *Meaning and Method* (New York: Cambridge, 1990), pp. 31-61; "Names without Bearers," *The Philosophical Review*, CI (1994): 1-39; *Sense, Reference, and Philosophy* (forthcoming).

(D) The sense of an expression is that aspect of its structure which is responsible for its sense properties and relations, that is, having a sense (meaningfulness), sameness of sense (synonymy), multiplicity of sense (ambiguity), repetition of sense (redundancy), opposition of sense (antonymy), and so on.

What makes (D) a non-Fregean definition of sense is that (D) makes sense the determiner of *sense* properties and relations rather than the determiner of *referential* properties and relations.

The pay-off for switching to (D) is that senses no longer have to contain the substantive properties that trigger Millian counterexamples. On the classical description theory, senses, being reference determiners, must contain a cluster of substantive properties which suffices to determine reference. Otherwise, they do not qualify as bona fide senses. With (D), it becomes possible to construct a version of the description theory on which the sense of a name contains no substantive properties. None is needed because we no longer have to ensure that the sense of a name determines its referential properties and relations. Thus, (D) enables us to formulate the one description theory which avoids Millian counterexamples.

This theory, the *pure metalinguistic theory* (PMT for short), says that the sense of a name contains only the nominal property of being the name of its bearer(s). On PMT, the sense of a proper name '*N*' is an instance of (1).

(1) *the thing which is a bearer of 'N'*

where "*x* is a bearer of '*N*'" is understood as expressing the relation between an individual and a name it bears. PMT can take this view because (D) requires that decisions about whether a term has a sense, and about what sense it has, are to be based on an inference to the best explanation of the term's sense properties and relations. For example, we take the sense *the thing which is a bearer of 'John Stuart Mill'* as the sense of the name 'John Stuart Mill' not because doing so enables us to determine its referent but because doing so enables us to explain inter alia why (2) analytically entails (3).

(2) John Stuart Mill wrote *A System of Logic*
(3) A bearer of 'John Stuart Mill' wrote *A System of Logic*

Similarly, taking 'Socrates' to mean *the thing which is a bearer of 'Socrates'* enables us to explain why (4) is analytic, or, as John Locke put it, trifling.

(4) Socrates is a bearer of 'Socrates'

There is, of course, the question of whether purity—the absence of substantive properties in the senses of names—is enough to guarantee the absence of Millian counterexamples. It is enough if the nominal properties appearing in instances of (1) do not themselves generate such counterexamples. Shortly, I shall explain why they do not, and, subsequently, describe the bearing of these considerations on a challenging thesis of Kripke's about epistemology and modality.

Even the most sophisticated Fregean theories of names (for example, Searle's) face Millian modal problems, and, by the same token, even the most sophisticated Millian theories of names (for example, Kripke's) face Fregean semantic problems. PMT faces none of those problems because it has none of the features responsible for them.

Millian modal problems arise for Fregean theories because Fregean senses of names contain substantive properties. The need for Fregean theories to say that names have a sense, which arises from the Fregean puzzles that motivate the introduction of senses, translates automatically, on such theories, into the need to say that names have a Fregean sense. So, the sense of a name is taken to have the substantive properties required for reference determination, and, as a consequence, there are modal counterexamples to the description alleged to express the sense of the name. Given (D), we can have senses that are instances of (1), and, hence, we can have Fregean solutions to Fregean puzzles without Fregean senses.

To be sure, Kripke thinks modal counterexamples arise for metalinguistic description theories, too, but, unlike PMT, the metalinguistic description theories for which they arise are impure in that they supplement the metalinguistic relation in the senses of names with substantive properties. William Kneale[9] claimed that 'Socrates' means *the individual called 'Socrates'*, since "it is obviously trifling to tell [someone] that Socrates was called Socrates" (*ibid.*, pp. 629-30). Kripke first objects to Kneale's theory because "it is dubious that the Greeks *did* call him 'Socrates' "[10] This criticism does not apply to our explanation of the analyticity of (4) because "*x* is a bearer of '*N*'" and "*x* is called '*N*'" are different relations. The former but not the latter is the converse of "'*N*' is a name of *x*." The latter means something like 'is addressed as', and, accordingly, whereas the *y*-place of the bearer relation takes only proper nouns, the *y*-place of the relation "*x* is called *y*" takes common nouns as well—for example, "Smith was called a thief and a dastardly murderer " Thus, while Kripke is right to say that sentences like 'Socrates was called 'Socrates" are not trifling or analytic, and might be false, (4) is analytic and hence cannot be false (*ibid*) Kripke's counterexamples to Kneale's impure metalinguistic description theory do not carry over to PMT[11]

---

[9] "Modality, De Dicto and De Re," in Ernest Nagel, Patrick Suppes, and Alfred Tarski, eds., *Logic, Methodology, and Philosophy of Science* (Stanford University Press, 1962), pp. 622-33.

[10] *Naming and Necessity*, p. 69.

Kripke's "main" objection to Kneale's metalinguistic description theories is that "[such a theory] would give a clear violation of the noncircularity condition " As Kripke puts it

> According to Kneale, the description [for the name 'Socrates'] is 'the man called 'Socrates' ' And here, (presumably, since this is supposed to so trifling!) it tells us nothing at all Taking it in this way it seems to be no theory of reference at all We ask, "To whom does he refer by 'Socrates'?" And then the answer is given, "Well, he refers to the man to whom he refers "[12]

Kripke is right that a theory on which the senses of names are genuinely trifling is "no theory of reference at all " But what of it? Not all description theories are intended to be theories of reference Since PMT is intended to be a theory solely of the senses of names, the sense that PMT provides for a name like 'Socrates' is required to determine only the sense properties and relations of the name

Fregean semantic problems arise for Millian theories because those theories deny that proper names have senses at all The need to avoid Mill-style counterexamples, which arises when the senses of names contain substantive properties, translates, on the Millian's equation of the senses with Fregean senses, into the need to avoid senses altogether But, given (D), the intensionalist claim that names have a sense and the Fregean claim that names have a reference-determining sense come apart Hence, the Millian can be right to deny that names have a reference-determining sense, but wrong to deny that they have a sense Thus, (D) provides us with the option of saying that names have a sense but that their sense determines no more than their *sense* properties and relations.

---

[11] A further point in connection with Kripke's parody is that we can also show, on the basis of (D), that 'horse' does not mean *thing named 'horse'* or *thing bearing the name 'horse'* Aside from the issue Kripke (*Naming and Necessity*, p 69) raises, the point is important for our classification of nouns at the beginning of section III

If common nouns could be so represented, then, on the construal of 'Boys are bearers of 'Boy' ' where the subject is the (plural) common noun 'boys', the sentence would be analytic in the manner of (4)

But, on this construal, the utterances of the sentence are false Manny, Moe, and Jack are named 'Manny', 'Moe', and 'Jack', not 'Boy' Actually, the sentence is ambiguous But it is, on the other construal—the one on which its subject is the proper noun 'Boy' and the sentence is thereby parallel to 'Smiths are bearers of 'Smith' '—that 'Boys are bearers of 'Boy' ' is analytic For example, Tarzan's son is named 'Boy', see the 1939 MGM film *Tarzan Finds a Son*

[12] *Naming and Necessity*, p 70

Such thin senses are all that is necessary to solve the puzzles that Millian theories fail to solve. Once Millianism is adopted, Frege's[13] puzzle about identity sentences with names resurfaces because all we are left with to distinguish sentences like (5a) and (5b) are referential and orthographic differences between the proper names in them.

(5a) Hesperus is Hesperus.
(5b) Hesperus is Phosphorus.

Neither reference nor orthography, as Frege himself pointed out, can account for the fact that sentences like (5a) "hold *a priori* and, according to Kant, are analytic," while sentences like (5b) "contain very valuable extension[s] of our knowledge" (*ibid.*, p. 56). But if we have senses for names, even ones as thin as on PMT, we can account for the difference in meaning between sentences like (5a) and (5b). Taking 'Hesperus' and 'Phosphorus' to have senses that are instances of (1) is taking these names to have different senses, which is all that is required to explain why (5a) is analytic and (5b) is not.

Similarly, taking 'Londres' and 'London' to have senses that are instances of (1) is taking them to have different senses, and this is all that is required to solve Kripke's Pierre puzzle.[14] Since Pierre's beliefs are now *the thing which is a bearer of 'Londres' is pretty* and *the thing which is a bearer of 'London' is not pretty*, Pierre's beliefs and statements concerning the city he read about and the city in which he is currently living are different, and, hence, not inconsistent, and there is no puzzle.

Scott Soames[15] says: "Clearly, [Pierre] has said both that London is pretty and that London is not pretty." (*ibid.*, p. 407). But Pierre never believed or said that London is pretty. One can, of course, say that Pierre believes or said of London that it is pretty, but this is not saying that he believes or said that London is pretty. If Oedipus believed and said that Jocasta is his wife, then Oedipus believed and said of his mother that she is his wife, but this is not to say that he believed and said that his mother is his wife. What Pierre believed and said is that *the thing which is a bearer of 'Londres' is pretty*, something quite consistent with believing and saying that *the thing which is a bearer*

---

[13] "On Sense and Reference" in *Translations from the Philosophical Writings of Gottlob Frege*, Max Black and Peter Geach, eds. (Oxford: Blackwell, 1952), pp. 56-57.

[14] "A Puzzle about Belief," in Nathan Salmon and Scott Soames, eds., *Propositions and Attitudes* (New York: Oxford, 1988), pp. 102-48.

[15] "Direct Reference and Propositional Attitudes," in *Themes from Kaplan* (New York: Oxford, 1989).

of 'London' is not pretty PMT thus explains why faultless reasoning from acceptable empirical evidence does not land Pierre in contradiction Moreover, that they should not land him in contradiction is surely everyone's intuition

Frege's problem about the meaningfulness of expressions and sentences containing bearerless names also resurfaces when Millianism is adopted[16] The most acute form of this problem, as I argued elsewhere,[17] is the case of true, hence meaningful, sentences like (5c)

(5c) Santa Claus does not exist

Since Millianism claims that bearerless names like 'Santa Claus' are also senseless, such names have nothing to contribute to the compositional meanings of sentences Hence, Millians must deny the meaningfulness of perfectly meaningful existence sentences like (5c) These sentences are counterexamples to Millianism In contrast, on PMT, bearerless names like 'Santa Claus' have a sense, and, hence, have something to contribute to the compositional meanings of sentences Accordingly, existence sentences like (5c) have a meaning

Further, we should indicate briefly how PMT handles Kripke's Paderewski case[18] Someone who realizes that the politician and the musician are one and the same person might express the realization by saying "Paderewski the politician is Paderewski the musician " In this case, the dummy term 'thing' in the sense of the first occurrence of 'Paderewski' would get filled with the sense of the apposition 'the politician', and the dummy term 'thing' in the sense of the second occurrence of 'Paderewski' would get filled with the sense of the apposition 'the musician' Thus, the sentence would have the sense *the politician who is a bearer of 'Paderewski' is the musician who is a bearer of 'Paderewski'* Since filling of the dummy terms takes place in the process of forming compositional senses for 'Paderewski the politician' and 'Paderewski the musician', the fact that the senses of these noun phrases contain substantive properties has no bearing on the status of PMT's claim that a proper noun like 'Paderewski' contains no substantive properties (1) asserts nothing about the senses of constituents in which proper names occur as proper parts

In a case where the speaker says simply, "Paderewski is Paderewski," it must be clear in the context that the statement is the denial of a prior presumption that the Paderewskis are different people PMT provides the sentence meaning required for such a prior presumption. The indefinite quantifier in the sense of 'Paderewski' provides the possibility of disjoint reference, the definite quantifier and the term provide the possibility of linking the occurrences of 'Paderewski' with background information (for example, about occupation) that makes them contextually definite in ways suggesting disjoint reference to someone with certain beliefs (for example, about occupation).

---

[16] "On Sense and Reference," p 58

[17] "Names without Bearers," pp 1-4, and p 34, footnote 45

[18] "A Puzzle about Belief," pp 102-48

Eliminating substantive properties from the senses of proper names eliminates modal counterexamples just in case the nominal properties in instances of (1) do not give rise to such counterexamples. David Braun[19] contends that they do; he says that PMT gets "the wrong modal result" in the case of an analytic sentence like (4) because its analyticity commits PMT to saying that Socrates bears 'Socrates' in all worlds where he exists, but, in fact, the name something bears is only one of its contingent properties (*ibid.*, pp. 559-60). Braun's criticism rests on a mistake about PMT, but the mistake might not have been made had my earlier discussions of names been clearer about why it is that taking sentences like (4) to be analytic does not commit PMT to claiming that something's bearing the name it bears is one of its necessary properties. Those discussions contain the principle which explains why there is no such commitment, but they do not use it to provide the explanation.[20]

Braun's mistake is to equate the intensionalism within which PMT is formulated with Frege/Carnap intensionalism. To be sure, within Frege/Carnap intensionalism, analytic sentences express necessary (or weakly necessary) truths, and, hence, a metalinguistic description theory formulated within it would get "the wrong modal result."[21] But the intensionalism in which PMT is formulated differs from Frege/Carnap intensionalism in how it understands analyticity as well as in how it understands sense. Not only are senses not reference determiners, analytic sentences are not necessary (or weakly necessary) truths. In *Semantic Theory*,[22] I defined a sense of a simple sentence—one having the form of a predicate $P(x_1, \ldots, x_n)$ with terms $t_1, \ldots, t_n$ occupying its argument places—as being analytic if there is a term t_i, $t_1 \leq t_i \leq t_n$, that contains $P(t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$. The difference between containment here and containment in Frege's definition is that the former is literal "beams in a house" containment while the latter is figurative "plants in their seeds" containment [23] Our analyticity is mereological structure, Frege's is logical structure

---

[19] "Katz on Names without Bearers," *The Philosophical Review*, CIV, 4 (1995): 559-60.

[20] Cf. my "Has the Description Theory of Names Been Refuted?" pp. 37-46.

[21] Kripke, "Identity and Necessity," in Milton K. Munitz, ed., *Identity and Individuation* (New York: University Press, 1971), pp. 135-64, here p. 137.

[22] New York: Harper and Row, 1972, pp. 171-97.

Frege's analyticity is provability from laws of logic plus definitions (*ibid*, pp 3-4) An analytic statement, to use W V Quine's[24] formulation of Fregean analyticity, is a logical truth in which we have put "synonyms for synonyms" (*ibid*, pp 22-23) Since Fregean analyticity is thus a species of logical truth, the referential interpretation of Fregean analyticity is fixed by the referential interpretation of logical truth Not so with our analyticity Since the mereological structure of analytic sentences does not make them instances of logical truths, it is open to us to provide a different principle of referential interpretation for analytic sentences In earlier publications,[25] I proposed the principle that analytic sentences are "secured against falsehood" That is to say, if a literal use of an analytic sentence makes a statement, then it is a true one For, if the containing term refers, then, in virtue of the containment, it refers to a state of affairs in which the statement is true, and if the term does not refer, the analytic sentence makes no statement

Security against falsehood does not imply necessity or weak necessity If I use a token of (4), the token must, the world being as it is, make a true statement, but the statement that it makes is contingent The statement is about someone who might not have been a bearer of 'Socrates' Since PMT is not committed to claiming that the statements we make with (4) are true in all worlds where Socrates exists, the existence of Socrates in a world where he does not bear the name 'Socrates' is no counterexample to PMT PMT does not get "the wrong modal result"[26]

---

[23] *Foundations of Arithmetic*, p 101

[24] "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge Harvard, 1953), pp 20-46

[25] *Semantic Theory*, p 179, "A Proper Theory of Names," pp 37-38, "The Neo-Classical Theory of Reference," p 115

[26] I want to thank David Pitt for discussions which helped clarify my thinking on some of these issues For further discussion of Braun's criticisms, see Bradley Armour-Garb and Katz, "What's in a Name" (forthcoming)

Note that we can make the referential interpretation of analytic entailments like that of (3) by (2) correspond to the referential interpretation of analyticity if we take the former to be preservation of security against falsehood Note also that there is a class of analytic sentences, including sentences like 'Bachelors are unmarried' or 'Cats are animals', which, in addition to being secured against falsehood, are weakly necessary truths—see my "Analyticity, Necessity, and the Epistemology of Semantics," *Philosophy and Phenomenological Research*, LVII, 1 (1997) 1-28, and *Sense, Reference, and Philosophy*

The conclusion is that PMT avoids the modal counterexamples that Fregean descriptivism does not avoid and solves Frege's puzzles about sense that neo-Millianism does not solve Note that this conclusion makes no claim that PMT solves all the puzzles that come down to us from Frege's "On Sense and Reference " Not all of them are puzzles about sense, and PMT is not obligated to solve those which are not about sense In particular, Frege's puzzle about the substitution of coreferential names in opaque contexts, though commonly lumped together with the puzzles about sense, is not such a puzzle [27] To be sure, it seems to go with Frege's puzzles about identity sentences with names and bearerless names, particularly as he presents them all as problems solvable on the basis of his notion of sense But, seen from the perspective of (D), the puzzle about substitution does not belong with the puzzles about identity sentences with names and sentences with bearerless names The latter are problems about how to explain sense properties and relations of sentences, respectively, analyticity and meaningfulness, but the former is not The puzzle about substitution is a problem about the conditions for valid substitutional inference—that is, it is about the truth of statements Hence, it belongs to the theory of reference It is thus outside a theory of sense based on (D) [28]

We come now to the issue about which, as I shall argue below, PMT and neo-Millianism are both wrong PMT and neo-Millianism say that (T) is false but it turns out to be true

(T) Names can have a sense consisting of substantive properties (as the classical description theory claims)

The truth of (T) is bound at this point to appear surprising because of the strength of Millian reasoning when deployed with the full range of Kripke's counterexamples But the truth of (T) in fact has nothing to do with the logical status of Millian counterfactual arguments, which do refute classical descriptivism Rather, it has to do with the semantics of a special class of poorly understood expressions, ones which I shall refer to as *improper names*

---

[27] "On Sense and Reference," pp 64-67

[28] The point here is of more than historical interest It undercuts the assumption of neo-Millians like Soames (*op cit*) that the issue about the meaning of proper names to which they believe direct reference is the definitive answer must be dealt with in the same theory that treats the statements that uses of propositional attitude sentences make in context The point shows that considerations from this theory which have been used to motivate the direct reference story about proper names does not in fact do so

So, here are the theses of this paper (T) is true, and, accordingly, neo-Millianism and PMT require revision The revision in the case of PMT is minor, and the minor revision is the end of it Essentially, the same minor revision in the case of neo-Millianism seems at first also to be the end of it, but things turn out otherwise Furthermore, examination of the largely neglected question of how names obtain their unique reference shows that such reference, even on Kripke's theory, requires descriptions[29] This brings the issue between neo-Millianism and PMT down to Millianism Then I present an argument to refute Millianism The bottom line is that the revised PMT is our only tenable theory of names

II MULTIPLE BEARERHOOD

In this section, I want to show that the meaningfulness of proper names is the only issue between neo-Millianism and PMT It looks at first as if the issue of whether descriptions mediate the reference of names also divides the two theories, especially insofar as neo-Millians advertise their theory as description free This, however, is an appearance created, on the one hand, by a proposal of Kripke's that neo-Millians must reject, and, on the other, by not recognizing how much of his causal picture of naming non-Fregean intensionalists can accept

II 1 Linguistic evidence The vast majority of proper names in natural language have multiple bearers, yet the vast majority of their tokens have unique reference Since classical descriptivists treat a proper noun as if it were semantically a common noun, they try to reconcile multiple bearerhood with unique reference by saying that proper names are ambiguous, explaining the unique reference of the tokens of "ambiguous" proper names in terms of disambiguation[30] But proper names are not ambiguous They are only referentially equivocal They have multiple bearers, not multiple senses If I use a token of (6) in speaking to philosophers

(6) Nagel has taught in New York City

their problem is not how to choose between two senses of a word (the name 'Nagel'), as, for example, it would be if I had used a token of (7)

(7) Let's meet at the bank

but how to choose between the two philosophers Ernest and Thomas (the bearers of 'Nagel') We disambiguate a common noun by citing a difference between its senses, but we cannot do this with proper nouns What sense difference disambiguates the proper noun 'Nagel' in the way that the sense difference *slope* and *financial establishment* disambiguates the common noun 'bank'?

---

[29] *Naming and Necessity*, p 96

[30] For example, Russell, p 525

Millians, too, have a problem with multiple bearerhood According to Mill, a name is a meaningless mark that distinguishes the object(s) to which it refers on the basis of its sound or orthography Recall his example of the robber in the *Arabian Nights* who put a chalk mark on a house for purposes of identification, but

> Morgiana chalked all the other houses in a similar manner, and defeated the scheme how? simply by obliterating the difference of appearance between that house and the others The chalk was there, but it no longer served the purpose of a distinctive mark [31]

Mill's account fails to explain how speakers of a natural language perform the identifying function in their use of names Since all the John Smiths are bearers of the name 'John Smith', "the difference of appearance" on which Mill's account of the identifying function of proper names depends is not there For Mill, speakers are like the robbers that Morgiana foiled

Neo-Millians who follow Mill in taking the distinguishing function of a proper name to be accomplished on the basis of sound or orthography also have a problem with multiple bearerhood Since the sound and orthography of the name 'John Smith' is the same from one John Smith to another, 'John Smith' does not serve as a distinguishing mark among the many John Smiths In the best of all possible worlds for the neo-Millian, names would mark their bearers in the same precise way as numbers assigned to prisoners, but, in the actual world, the sound and orthography of names in natural languages is not a notation that keeps track of their bearers like prison numbers

Hence, there is considerable pressure on neo-Millians to go beyond Mill and concoct a notation on which the name of one John Smith is somehow a different mark from that of another Kripke succumbs to this pressure, proposing that we

> adopt a terminology according to which uses of phonetically the same sounds to name distinct objects count as distinct names [32]

Kripke says this terminology is "analogous to the practice of calling homonyms distinct 'words'" (*ibid.*) This, however, is patently false Homonyms, such as 'bear' and 'bare', differ in etymology, (typically) orthography, syntax, and meaning. Kripke's distinct names, such as 'Smith' used to name one person and 'Smith' used to name another, do not differ in etymology, orthography, syntax, or meaning (since both names 'Smith', being names, are supposed to have zero meaning).

---

[31] *A System of Logic*, I, 2, 5, 36

[32] *Naming and Necessity*, p 8

Kripke's proposal might be, and sometimes has been, construed as claiming that occurrences of a name like 'Smith' have unpronounced subscripts that distinguish occurrences of 'Smith' in terms of the different bearers of the name. Since this claim constitutes a substantive hypothesis about the language, it has to be backed up with substantive linguistic evidence, particularly, as accepting the claim would sacrifice a long-standing linguistic principle, namely, Noam Chomsky's autonomy of syntax.[33] Linguistic methodology requires grammatical evidence for positing underlying grammatical forms, but there is no such evidence for claiming that unpronounced subscripts are part of the grammatical form of proper names.[34]

Further, as Kripke himself concedes, his terminology "does not agree with the most common usage."[35] Shortly, we shall examine the conflict with the linguistic fact, but, before we do, we should look at the two considerations that Kripke offers to mitigate the concession. First, he suggests that his terminology is appropriate for the question 'How many names are there in this telephone book?' But to suppose that the answer to this question is got by counting each occurrence of a name as a distinct name, rather than by counting each simply as a distinct occurrence of the same name, begs the question. Second, Kripke says that his proposal has "a great deal to recommend it for theoretical purposes" (*ibid.*). He does not elaborate, but, presumably, the purposes in question are not providing neo-Millianism with different names for different bearers, since, then, Kripke would be arguing from his own theory. The only other thing that suggests itself is the purposes the terminology might serve in logic—for example, the purposes it serves in Kripke's quantified modal logic.[36] But the usefulness of a formalism for an artificial calculus implies nothing, one way or the other, about its usefulness for a grammar of a natural language.

---

[33] That is, the principle that syntactic rules cannot refer to pragmatic, phonological, or semantic information—*Essays on Form and Interpretation* (Amsterdam: North-Holland, 1977), pp. 36-59.

[34] For a discussion of the kind of evidence required for positing underlying phonological structure, see M. J. Kenstowicz, *Phonology in Generative Grammar* (Cambridge: Blackwell, 1994), pp. 107-14.

[35] *Naming and Necessity*, p. 8.

[36] Kripke, "Semantical Considerations in Modal Logic," *Acta Philosophica Fennica*, xvi (1963): 83-94.

Whereas classical descriptivists proliferate senses for referentially equivocal names, Kripke proliferates names themselves [37] Kripke's proliferation, however, is not the ordinary sin against parsimony in which the offending proposal accounts for the same evidence less economically Here the proposal which offends against simplicity is also at variance with the evidence over a wide range of linguistic structures

The evidence clearly shows that the proper noun 'John Smith' is a single name with many bearers We quite naturally say such things as (8) and (9)

(8) The fifty John Smiths living in New York are Yankee fans
(9) I know one John Smith, but there must be more in America

But, on Kripke's proposal, this is deviant Most strikingly, his proposal denies that two John Smiths are speaking the literal truth when they make a claim like (10) or (11)

(10) We have the same name
(11) 'John Smith' is my name, too

It also affirms that two John Smiths are speaking the literal truth when they say (12)

(12) We have different names

---

[37] Underlying both proliferations is a conflation of a feature of the structure of language—the possibility of proper names having multiple bearers—with a feature of the use of language—the fact that referentially equivocal types have contextually definite tokens The conflation becomes evident if we ask how all the distinct Kripkean names 'John Smith' are to be individuated Homonyms are individuated in terms of linguistic factors concerning etymology, orthography, and meaning, but such factors do not apply in the case of Kripkean names We have to appeal to extralinguistic considerations, in particular, the different bearers in each case This is because specifying those names requires us to collect uses of the proper name, tokens of 'John Smith', into equivalence classes on the basis of something like the idea behind Russellian propositions Each equivalence class of utterances representing a Smith "homonym" contains every token whose referent is a particular Smith plus the Smith who is their referent

The referent must be included in the equivalence class, since otherwise there will be many cases where different names are associated with the same equivalence class Note cases like the drill sergeant who uses an utterance of 'Smiths front and center!' to refer to several Smiths Since the sergeant's utterance will appear in more than one equivalence class, the equivalence class for each of these Smiths has to include the relevant Smith Since, on Kripke's proposal, the many names 'John Smith' are each different words in spite of the fact that they are linguistically indistinguishable, having the same etymology, phonology, orthography, syntax, and (ex hypothesi null) sense, the names must include their referents as a linguistically distinguishing feature The inclusion of a 'Smith' referent as a part in the constitution of each distinct 'Smith' name makes the language/language-use conflation evident

Also, Kripke's proposal says that two John Smiths are making the same claim about their name in saying (12) as the claim that a John Smith and a Mary Potter are making about their names in saying (12). Further, it hard to see how to make sense of the definite article in sentences like (13) without taking the name to have more than one bearer.

(13) *The* Albert Einstein lived in Princeton.

Still further, one John Smith can "prove" to another John Smith that his name is not 'John Smith'. Premise 1, "My name is 'John Smith'", premise 2, "Kripke's proposal is true", conclusion, "Your name is not 'John Smith'."

Moreover, since Kripke's proposal is a linguistic thesis, accepting it commits one to saying that the English word 'namesake' in a sentence like (14) is a necessarily vacuous term.

(14) My name is Mary Smith, too, so I'm her namesake.

It is no more possible for someone to have a namesake than to have a male sister.

Another aspect of the situation concerns name change. If Monica Lewinsky married Kenneth Starr, she would have the choice of keeping her surname or changing it to her husband's, but, on Kripke's proposal that "distinctness of the referents will be a *sufficient* condition for distinctness of names," the choice does not exist.[38] On his proposal, even if Monica Lewinsky does take the name 'Starr', she has not changed her name 'Lewinsky' to her husband's name.

This case is run together with the case in which Brenda Starr marries Kenneth Starr. Even though Kripke's proposal would give the right answer for the latter case, it gives it for the wrong reason. The reason Brenda Starr has no choice about whether to change her name when she marries Kenneth Starr is not that distinctness of referent makes it logically impossible to have another's name, but that, since her husband's name is 'Starr', too, it makes no sense to talk about her changing her name. There is no such thing as *changing* the name 'Starr' to the name 'Starr'.

Yet another aspect of the problem concerns family names. On Kripke's proposal, there are no family names in the customary sense that all the members of the Smith family have the same name 'Smith'. Even more counterintuitive, none of the members of the family can really have the family name 'Smith', since its bearer ought to be a family, hence a collection, not an individual. How benighted of the

---

[38] *Naming and Necessity*, p. 8, footnote 9.

Smith family to be concerned about passing down the good name 'Smith' or be worried about Raymond, the black sheep of the family, besmirching the family name!

Furthermore, if the name of the son, George Smith, were, as it is on Kripke's proposal, a different name from that of the father after whom he was named, there would be no point to the 'Junior' associated with the son's name. On Kripke's proposal, the adjective 'Junior' in a sentence like (15) would be redundant or unnecessary, but it is not.

(15) I'd like you to meet George Smith, Junior.

Note, further, that even a neo-Millian who rejects Kripke's proposal has no way to explain how the adjective 'Junior' enters compositionally into the sense of (15). Using the quantifier structure represented in (1), PMT can say the adjective enters by way of attributing the property of being the son of another bearer of '*N*' to the bearer of the '*N*' functioning as the object of the verb. PMT thus explains why (15) implies that there are or were at least two George Smiths.

Finally, lexicographers say such things as that the proper noun 'Smith' is related to the Old Norse 'Smitr', the Danish and Swedish 'Smed', and the Gothic 'Smitha'. They take such statements to be about the etymology of *the* modern English word 'Smith'. Surely, they are right that there is but one word 'Smith' in English in just the (type) sense in which C. S. Peirce held that there is one word 'word' in English. Now, each Smith can, with equal justice, proudly say, "It's my name that the lexicographers find so interesting." Moreover, it makes no sense for two Smiths to argue about whose name is the one being talked about when lexicographers discuss the relation of 'Smith' to the Old Norse 'Smitr'. But, on Kripke's proposal, it does. Further, insofar as on that proposal, there is no single name 'Smith' in the type sense, the proposal implies, absurdly, that the lexicographers are talking about nothing.

Kripke's way of putting his concession—that the terminology "does not agree with the most common usage"—makes it look as if nothing more is involved than deviation from certain superficial features of linguistic custom. In fact, from the standpoint of the linguistic evidence just cited, Kripke's proposal about when to count phonological types and tokens as distinct proper names represents a throwback to the constructionalist programs of earlier philosophers like Frege, Carnap, and Gustav Bergmann to reform natural language. Kripke's proposal is a clear case of what Wittgenstein called "purify[ing], sublim[ing] the signs themselves," on all fours with Frege's proposal to reform the language by assigning referentially vacuous expressions an arbitrary referent or eliminating them entirely[39] It is not easy to see why there should be so much more tolerance for Kripke's subliming the language in the case of names with two or more bearers than there is for Frege's subliming the language in the case of names with zero bearers

II 2 Philosophical consequences Names are not like the numbers assigned to prisoners Neither for proper nouns nor for their tokens is there a phonologically or orthographically distinct name for each individual bearing the name Nor are names ambiguous like common nouns Multiple bearerhood contradicts both Kripke's proposal and the classical description theory PMT, in contrast, is consistent with multiple bearerhood, but, more than this, it predicts the possibility that names have multiple bearers because (1) has an indefinite quantifier associated with the bearer place of the name relation Neo-Millians cannot, in principle, predict this possibility as a consequence of the grammar of proper nouns, since, in not allowing names to have senses, there is no place for an indefinite quantifier that introduces the possibility of more than one bearer

To square their theory of names with the linguistic facts, neo-Millians must reject Kripke's proposal But rejecting the proposal opens a gap in the neo-Millian account of how a speaker achieves unique reference in the use of a name, and, to close it, neo-Millians have to invoke descriptions, thereby sacrificing their claim to have a description-free theory

Consider Kripke's theory of name transmission

> When the name is 'passed from link to link', the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it[40]

In the case of names with multiple bearers, we receive the same name from different teachers who use it with different reference Each and every such transmission is the last link in a distinct chain leading back to a different baptism For a popular name like 'Smith', we are awash in causal chains Each of them involves the same name, 'Smith', but a different Smith Hence, intentions of speakers to use a name with the same reference as the people from whom they acquired it do not suffice to achieve unique reference

On the one hand, there is a communication problem Not being mind readers, the audience cannot know to which bearer of a referentially equivocal name the speaker is referring unless the speaker provides an individuating description On the other hand, there is a reference problem Without appropriate substantive content as part of the speaker's referential intention, he or she cannot use the referentially equivocal name to refer to the object named in a particular one of the baptisms to which the various chains involving the name lead back Big Louie has not referred to anyone if, believing that there are a number of Corleones in the Mob but having no identifying descriptions, he gives the police the name 'Corleone'

---

[39] Wittgenstein, § 94, Frege, *Translations from the Philosophical Writings of Gottlob Frege*, p 70

[40] *Naming and Necessity*, p 96

Since neo-Millian theory does not rely on descriptions to identify the intended bearer, the theory is subject to essentially the same criticism Kripke makes of Kneale's theory Paraphrasing we ask "To whom does he refer by 'Smith'?" And the answer is given "Well, he refers to the object named in the baptism initiating one of the chains for the name 'Smith' " Here Kripke's objection to Kneale's theory comes home to roost

> If this were all there was to the meaning of a proper name, then no reference would get off the ground at all (*ibid.*, p 70)

Since neo-Millian theory is a theory of reference, its claim that the application of names does not depend on descriptions is a casualty of multiple bearerhood Such theories can no longer be advertised as description free

But, although the misleading advertising has to stop, the product does not have to come off the shelves A new and improved version of neo-Millianism can be sold as a purely referential account of names in which descriptions enter as part of the use of names, functioning to distinguish the intended bearer from other contextually salient bearers This is still a version of Millianism, since the neo-Millian is not forced to make the defeating concession that descriptions enter as senses of names

But when a neo-Millian account of names is revised in this way, descriptions play the same role on its account of naming that they play on PMT's account[41] Hence, the alleged issue between neo-

---

[41] The role of descriptions in PMT is established by the definite quantifier in (1) associated with the head of the construction It introduces the linguistic requirement of uniqueness for the reference of names, namely, that literal uses of (singular) names are limited to one bearer per use Given the information from the indefinite quantifier, speakers know that their use of a name must distinguish their intended bearer from other bearers of the name—or at least those known in the context to be bearers When more than one bearer is known in the context and none of the bearers are salient, the speaker is required to supply descriptions which identify the bearer to which he or she intends to refer (Note the cuteness of someone using 'Hey, Sam' to get the attention of both Sams in the room)

Thus, on PMT, descriptions which ensure definite reference for proper names appear in the context rather than in the sense of the name This has two immediate

Millianism and PMT about the role of descriptions disappears Further, since nothing prevents PMT from incorporating Kripke's conception of baptism and name transmission as an interpretation of the bearer relation in (1), there need be no issue between neo-Millianism and PMT with respect to the bestowing and transmitting of names [42] Hence, there is only one issue Millianism

Since criticism of classical descriptivism has come from Millians, a fundamental objection to classical descriptivism has been missed, namely, that the descriptions classical descriptivists take to express the sense of a name are inadequate from a linguistic standpoint This is because descriptions like 'a pupil of Plato and a teacher of Alexander' or 'led the Israelites out of Egyptian captivity' prevent us from accounting for sense properties and relations of sentences containing proper names like 'Aristotle' and 'Moses' For example, sentences like 'Aristotle was a pupil of Plato and a teacher of Alexander' and 'Moses led the Israelites out of Egyptian captivity' are not trifling as they should be if the description their predicate expressed were actually the sense of their subject Frege took sense to be of no special concern so long as the description chosen for a name gets its reference right, and classical descriptivists have embraced Frege's "tolerant" attitude [43] But this attitude is not linguistically adequate—nor, for that matter, is it consistent with Frege's explicit appeal to Kantian analyticity in his argument for senses on the basis of identity sentences

A recent approach to the classical description theory based on the actuality operator at first looks as if it avoids this problem by rigidifying the reference of proper names directly with respect to their reference in the actual world [44] The idea is to use a rigidified description like 'actually John Smith' or 'the John Smith in the actual world' to fix reference But multiple bearerhood prevents rigidifying the reference of names in this way The problem is that there are too many John Smiths in the actual world Which one of them is *the* John

---

pay-offs First, unlike Kripke's account of names, PMT keeps the linguistic possibility of multiple reference separate from the extralinguistic mechanism which imposes the constraint that referentially equivocal types have contextually definite tokens Second, the description which determines the reference of a token in one context does not automatically constrain other tokens of the same name type in other contexts Accordingly, PMT encounters none of the counterfactual counterexamples that plague the classical description theory where the same description determines the referent of all the tokens of a name type

[42] See "Names without Bearers," pp 17-23

[43] *Translations from the Philosophical Writings of Gottlob Frege,* p 58§*

[44] Frank Jackson, "Reference and Description Revisited," in J E Tomberlin, ed , *Mind, Language, and Ontology Philosophical Perspectives,* Volume 12 (Cambridge Blackwell, 1998), pp 201-18

Smith? In order to prevent the multiple bearers of a name like 'John Smith' from raising this problem, it is necessary to adopt Kripke's proposal, but, as we have seen, the evidence about multiple bearerhood in natural language rules it out

III IMPROPER NAMES

Contrary to both PMT and neo-Millianism, (T) is true There are names in English with senses containing substantive properties Their existence has at times been noted, but they have not received the attention necessary for their significance to be properly understood This, I submit, is because the categories Proper Noun and Name have been wrongly taken to be coextensive with the category Proper Name So, let us begin with a better classification

Both proper and common nouns have senses, but senses of the former contain the metalinguistic predicate '*x* is a bearer of '*N*'', while the senses of the latter do not Proper nouns divide into proper names, such as (16)

(16) Dartmouth, Aristotle, Socrates, Moses, Godel,

and improper names, such as (17a) and (17b)

(17a) Jack the Ripper, Attila the Hun, Richard the Lion Hearted, Ivan the Terrible, Dennis the Menace, Mandrake the Magician,

(17b) Superman, Batman, Spiderman, Plastic Man, Wonder Woman,

The senses of improper names, unlike the senses of proper names, contain substantive properties in addition to the metalinguistic predicate '*x* is a bearer of '*N*''

Consider some of the evidence in favor of (T) Improper names contain meaningful syntactic constituents For example, the definite descriptions in (17a), 'the Ripper', 'the Hun', 'the Lion Hearted', and so on, seem to be as straightforwardly meaningful as any other modification Indeed, the improper nouns in (17a) initially strike us as frozen forms of compositionally meaningful modifications We can easily imagine that the definite description 'the Terrible' first occurred as a modifier distinguishing that Tsar Ivan from the other tsars of the same name

Another form of evidence that senses of improper names contain substantive properties is the recursive expandability of the definite descriptions in improper nouns like those in (17a) Not only does the sense of 'Dennis' compositionally combine with the sense of 'the' and 'menace' to form the sense of 'Dennis the Menace', but modification is possible for the nouns within such improper names—for example, we have forms like (18)

(18) Dennis the Elm Street Menace, Dennis the Elm Street and School Yard Menace,

Such recursive expansions are compositionally meaningful at each step, and each step differs in meaning from the previous one by just the sense of the modifier added at that step

Still further evidence for the meaningfulness of improper names comes from the fact that, unlike proper names, they translate into a foreign language Thus, the proper name 'Ludwig Wittgenstein' and the proper name 'Heinrich Himmler' are not rendered 'Louis Wittgenstein' and 'Henry Himmler' in English, but the improper name 'Richard the Lion Hearted' is rendered 'Richard Coeur de Lion' in French

The sense of the improper name 'Superman' contains a substantive property something like *man with superhuman powers* (for example, leaping tall buildings in a single bound and moving faster than a speeding bullet) The sense of 'Batman' contains a substantive property something like *man who uses bat symbolism in his superhero role* The sense of 'Plastic Man' contains a substantive property something like *man able to stretch his body into virtually any shape*

Having a different grammatical structure from the examples in (17a), improper names like those in (17b) are not expandable productively Whereas 'Ivan the Terrible' is a phrase containing a proper noun and a modifier, 'Batman' is a compound common noun like 'blackboard' The productivity in the latter cases is thus different It results in compound proper nouns, for example, 'Batmobile' with the derivative meaning *automobile designed with bat symbolism for Batman to use in his superhero role*

To show that improper names like those in (17b) are not proper names, let us imagine a superhero who uses gold symbolism in his superhero role and who is, accordingly, called 'Goldman' The meaningfulness of the improper name 'Goldman' is evident when we compare it to the orthographically identical proper name 'Goldman' Unlike the superhero, a bearer of the proper name 'Goldman' can be a woman and have no special relation to gold Further, the meaningfulness of improper names is shown by fact that the plural of the improper name is 'Goldmen', as in (19a)

(19a) Three Goldmen came to the rescue in Goldmobiles

while the plural of the proper name is 'Goldmans', as in (19b)

(19b) Three Goldmans came to visit us in Oldsmobiles

That is to say, the occurrence of 'man' in the improper name, in contrast to its occurrence in the proper name, undergoes the customary pluralization from 'man' to 'men' with the customary shift in meaning from *one* to *many*.[45]

The substantive properties in the sense of improper names constrain their reference in essentially the same way as the substantive properties in the sense of common nouns. Thus, in literal uses, the substantive properties in the sense of an improper name express necessary properties of its bearer. No one can literally be Superman if he has only ordinary human powers. Nor can the superhero be a female. Note that saying no woman can literally be Super*man* does not say that the improper name cannot be used to refer to a woman. Suppose the Kents mistakenly took Jar El, a female, to be male, and raised her as a male. Realizing that she can be a more successful superhero in the guise of a man, she maintains the fiction that she is one. On this scenario, the Metropolis citizens' use of 'Superman' ("Look in the sky it's Superman") to refer to a woman is a case of reference under a false description. In another kind of case, a cowardly Richard commanded his subjects to call him 'Richard the Lion Hearted'. In still another, a harmless Ivan, due to bad press, comes be to called 'Ivan the Terrible'. Such cases are also reference under a false description.

Because the senses of improper names involve substantive properties, improper names, in contrast to proper names, have substantive application conditions like common nouns. The proper name 'Ebeneezer Scrooge' applies to Scrooge even after he is frightened into becoming a model of Christian charity, but the improper name 'Dennis the Menace' does not apply to Dennis after he matures, changes his threatening ways, and becomes a harmless, well-intentioned adult. Technically, people might still call him 'Dennis the Menace', but only in a figurative way, in order, say, to remind him, meanspiritedly, of what he had once been like.

Things are more complicated with a case like 'Ivan the Terrible'. It seems reasonable to say that Ivan the Terrible bears this name even in old age when he is too feeble to strike fear in anyone. We might even say of him in his dotage "Ivan the Terrible is an old sweetie-pie." This shows that the assignment conditions for improper names do not have the same temporal structure as their application conditions. Putting this complication aside, the point is that, for Ivan to be Ivan the Terrible, he must have earned his bad reputation at some period during his life. Otherwise, the expression would be merely an ironic form like 'Li'l Abner'.

---

[45] The last point is due to Armour-Garb (in conversation).

Given the existence of improper names, (T) is true Thus, we must restrict (1) to be just a schema for senses of proper names and then add a separate schema for senses of improper names Hence, (1′)

(1′) (i) If '*N*' is a proper name, its sense is *the T which is a bearer of 'N'* ('*T*' is, in effect, the dummy term 'thing' in (1))
(ii) If '*N*' is an improper name, the sense of '*N*' is *the T* which is a bearer of 'N'*, where '*T**' is a term which, unlike the dummy term '*T*', contains substantive properties

In cases like (17a), the occurrence of *T* in the proper name component of an improper name like 'Dennis the Menace' becomes an occurrence of *T** in the compositional formation of a sense of the name the sense of the appositional phrase fleshes out the dummy term (see the above discussion of the Paderewski case) In cases like (17b), an occurrence of *T** is an inherent, that is, lexical, feature of the improper name

Kripke concedes that

> There may be some cases where the description picture is true, where some man really gives a name by going into the privacy of his room and saying that the referent is to be the unique thing with certain identifying properties 'Jack the Ripper' was a possible example which I gave [46]

Despite the fact that Kripke puts the concession as a possibility, it is quite clear that, with the grammatical distinctions available in *Naming and Necessity*, 'Jack the Ripper' is a proper name like any other, and, as a consequence, the expression poses a counterexample to the Millian view that names do not have a sense For, at the very least, 'Jack the Ripper' has the sense *causer of cuts or tears* Given that the literal referent of tokens of 'Jack the Ripper' is semantically required to be the causer, of cuts or tears, Kripke must concede that "the description picture is true" for some cases

Kripke also writes as if this concession is necessary only for one or another marginal case—as if the 'Jack the Ripper' case is the exception and cases like (16) the rule But the distinction here is not that between exception and rule but that between attributive uses of a description to fix the bearer of a name and referential uses (or ostension) The 'Jack the Ripper' case can serve as a model for a general strategy of introducing proper as well as improper names Could not a proper name 'Dartmouth' have had its reference fixed in the same manner as the improper name 'Jack the Ripper'? Surely, the British authorities could have gone into the privacy of their chambers and stipulated that the referent of 'Dartmouth' is the city, whichever it is, that is situated at the mouth of the river Dart If Dartmouth had been named in this way, then, on Kripke's own account, 'Dartmouth' would be synonymous with 'the city situated at the mouth of the river Dart' Hence, this proper name would have a sense

---

[46] *Naming and Necessity*, p 94

Gareth Evans[47] was the first, to my knowledge, to suggest that an attributive use of a description in a baptism can provide a proper name with substantive content His example was the name 'Julius', introduced by the stipulation that the name refers to the man, whoever it is, who invented the Zip Evans calls such names "descriptive names" because he takes them to have substantive content and thinks "it is not very plausible to deny that 'Julius' is a member of the same semantical category as other proper names" (*ibid.*, pp 31-32) The existence of an open set of cases of attributive uses of descriptions in baptisms begins to make it look as if Kripke's concession to "the description picture" is far more damaging than he led us to think

Evans's 'Julius' case, my 'Dartmouth' case, and perhaps even Kripke's 'Jack the Ripper' case are hypothetical cases, but there is reason to think that many actual cases of naming work the way they do The reason is that many situations will be like the one that is imagined in the 'Julius' and 'Jack the Ripper' cases There is a generalization these are cases where the baptizers have no direct access to the object to undergo baptism, where their only access to it is qua satisfier of a certain description This is the case in many theoretical baptisms, such as in Kripke's 'Neptune' case, and also in many mundane ones [48] Further, in some situations, the object does not yet exist, as when an inventor has the idea for a product and a marketing name but has yet to invent the bearer Hence, it seems reasonable to think that, in the course of history, there have been many baptisms based on attributively used descriptions [49]

---

[47] *The Varieties of Reference* (New York Oxford, 1982)

[48] *Naming and Necessity*, pp 79, 96

[49] Two things that Kripke says bear on this He says "First, the description used is not synonymous with the name it introduces but rather fixes its reference" (*Naming and Necessity*, p 96, footnote 42) This is unhelpful because, while what Kripke says may be true for perhaps the larger number of cases, the issue here is whether there are not also a significant number of cases where the description used to refer to the baptizee provides its name with a sense Kripke also says "Second, most cases of initial baptism are far from those which originally inspired the description theory" (*ibid.*) To be sure, this raises a problem for description theorists in cases of figures from the distant past where the relevant descriptions are unknown to contemporary speakers But the fact that classical descriptivists have their own troubles is no solace to neo-Millians For, even if we do not know specifically in which cases the bapitizer has employed a description attributively, it is enough to know that there are proper names which are synonymous

But, on closer inspection, names like Evans's 'Julius' turn out not to be counterexamples to the core Millian thesis. Evans says that "it is not very plausible to deny that 'Julius' is a member of the same semantical category as other proper names," but this is true only if neo-Millians are limited to the grammatical distinctions in *Naming and Necessity*. If they adopt the distinctions in our classification, neo-Millians can argue that a critic like Evans is not entitled to say anything stronger than that 'Julius' is a member of the category Proper Noun. Hence, neo-Millians have a response to Evans-style criticism: the proper noun 'Julius' is an abbreviation for an improper name. As Evans presents the case, the term 'Julius' is introduced on the basis of a speech act performed with some such sentence as "Let us call whoever invented the Zip 'Julius'."[50] The act requests us to use 'Julius' as short for the improper name 'Julius the Inventor of the Zip'. Since Evans's term 'Julius' is not a proper name, but, in effect, an improper name, it does not contradict the restricted Millian thesis that *proper names* have no sense.[51]

Hence, the mistake which led to Kripke's concession to "the description picture" was similar to our mistake and is similarly remedied. Neo-Millians should distinguish between proper names and improper names, and restrict the Millian thesis to the former. Nothing much is lost, since proper names were all along the only names of real concern to them.[52]

It looks, however, as if something important is lost nonetheless, namely, Kripke's claim that certain a priori truths are not necessary

---

with a description for neo-Millians to have to concede that some proper names have a sense.

[50] *The Varieties of Reference*, p. 31.

[51] The general strategy for the other cases in question is to model them on the preceding construal of the 'Julius' case, that is, to treat the name in cases where an attributive use of a description fixes the reference as an abbreviation of an improper name. Proper names, for neo-Millians, are proper nouns where a referential use of a description or an ostension in the baptismal ceremony fixes their reference. Proper names, the neo-Millian can continue to claim, have no sense.

On this treatment, attributively used descriptions do not introduce the possibility that a name like 'London' with multiple bearers is ambiguous. If the name 'London' has multiple bearers on the basis of attributive uses of *n* nonsynonymous descriptions, then we have ill-chosen abbreviations of *n* improper names.

[52] Even in the case of improper names, senses are not reference determiners. There could be two Dennis the Menaces. Or, there could be two Supermen, that is, two visitors from Krypton with the same superhuman powers, both of whom were raised by a kindly old couple named 'Kent', and so on, and both of whom go into the superhero business under the name 'Superman'. Hence, the senses of 'Dennis the Menace' and 'Superman' do not suffice to determine the definiteness of reference required in a use of a name.

truths.[53] To begin, note that truths expressed by sentences like 'Julius is the inventor of the Zip' count as necessary truths because their unabbreviated form is an analytic proposition like *Julius the Inventor of the Zip is the inventor of the Zip.* Thus, if, as Kripke assumes in *Naming and Necessity*, analytic propositions are instances of logical truths, he is wrong to deny that an a priori statement like (20) is a necessary truth.

> (20) Neptune [that is, the cause of such and such observed perturbations of the orbits of such and such planets] is the cause of such and such observed perturbations in the orbits of such and such planets.

If (20) is analytic in the manner of 'Julius is the inventor of the Zip', (20) is a necessary truth. Further, if Leverrier assigns the name 'Neptune' on the basis of the description 'the cause, whatever it is, of such and such observed perturbations in the orbit of such and such planets', it makes no sense for Leverrier to think that "some other object might have caused the perturbations in its place," as Kripke imagines in defending his claim about apriority and necessity. This is like thinking "Homer didn't write the *Iliad* and the *Odyssey*, it was some other ancient Greek by the same name." So, Kripke seems to have been wrong in claiming that a priori truths like (20) are not necessary truths.

On the assumption that analyticity is what Fregean intensionalism says it is, Kripke's claim is wrong. But if analyticity is what our non-Fregean intensionalism says it is, his claim is right. Since our notion of analyticity is not Frege's logical notion but rather the mereological notion described in section 1, an analytic sentence like (4) does not express a necessary truth in spite of the fact that we can know it is analytic a priori. On our intensionalism, the analyticity of (4) permits us to say no more than that our uses of it either make a true statement or no statement at all. In our world, (4) makes a true statement, but the statement is contingent because Socrates might not have been a bearer of 'Socrates'. Similarly, in our world, (20) makes a true statement, but the statement is contingent because Neptune might not be the cause of such and such perturbations in the orbits of such and such planets.[54]

IV. AN ARGUMENT AGAINST MILLIANISM

Although neo-Millians can still claim that proper names do not have a sense, neither Mill's argument nor any of the Kripkean variations can establish the claim. Such arguments are only arguments against Fregean senses. Millian arguments can (and do) decide whether proper names have Fregean senses, since that question falls within the domain of reference. They cannot decide whether proper names have senses, since this question falls within the domain of sense. I shall now give an argument that decides the question.[55]

---

[53] *Naming and Necessity*, p. 79, footnote 33.

[54] For, as Kripke points out, it could have "been knocked off its course one million years earlier"—*Naming and Necessity*, p. 79, footnote 33.

Neither Mill himself nor neo-Millians like Kripke, Putnam, and Keith Donnellan is an extensionalist.[56] Neo-Millians take common nouns, verbs, adjectives, adverbs, and so on to be meaningful, and, presumably, when presented with the evidence in the last section, they would agree, as Kripke already does in connection with the 'Jack the Ripper' case, that improper names, too, have a sense. Their denial that proper names have sense is not a consequence of an objection to senses per se, such as Quine's, but derives instead from what those philosophers take to be the special character of proper names. Since neo-Millians are thus intensionalist with respect to a large portion of the language, they require a theory of sense for the language as a whole. Without one, their semantics is incomplete in the manner of an account of gender that addresses one gender but not the other.

A theory of sense for the language must meet the criteria for linguistic theories. A linguistic theory, whether about phonology, syntax, or sense is justified on the basis of its success in explaining the relevant properties and relations of sentences. Accordingly, ascriptions of senses to expressions of the language must be based on the role the ascriptions play in explaining the sense properties and relations of sentences. If neo-Millians were to choose a theory of sense that fails to satisfy reasonable criteria for linguistic theories, their position would inherit the failure of the chosen theory of sense.

---

[55] My earlier attempts to answer this question were, for the most part, based on the strategy of providing a version of the description theory that escapes the problems of both the classical version and Millian theories. This strategy raised various tangential issues about semantic and pragmatic differences between the PMT and Millian theories that have obscured the fundamental issue between them. Here my strategy is to show that the issue is just Millianism and to argue directly against it.

[56] Putnam—"The Meaning of 'Meaning'," in Keith Gunderson, ed., *Language, Mind, and Knowledge* (Minneapolis: Minnesota UP, 1975), pp. 131-93—is up front about being a Fregean intensionalist. His criticism of Frege is only that senses as reference determiners leads to the externalist view that knowledge of meaning is not entirely in the head. Donnellan—"Necessity and Criteria," this JOURNAL, LIX, 22 (October 25, 1962): 647-58—is also an intensionalist, citing 'cat' and 'red' as examples of terms with senses. Donnellan changed his view on natural-kind terms, but that change does not entrain a change in view concerning color terms ('red'), artifact terms (for example, 'knife'), institutional terms (for example, 'bachelor'), linguistic terms (for example, 'anagram'), and mathematical terms (for example, 'prime'). Though more cagey, Kripke (*Naming and Necessity*, pp. 127-28), too, is an intensionalist, taking terms like 'foolish', 'fat', and 'yellow' as having senses.

A linguistic theory is in the business of explaining all the phenomena in its domain, that is, of explaining the linguistic properties and relations of all the expressions that have the appropriate linguistic properties and relations. Thus, theories of sense can show no favoritism with respect to explaining phenomena in the domain of sense. Just as theories of syntax can show no partiality toward the explanation of the syntactic properties and relations of one or another class of expressions, so theories of sense can show no partiality toward the explanation of the sense properties and relations of one or another class. Hence, neo-Millians are committed to a theory of sense that explains the sense properties and relations of all expressions and sentences of the language having such properties and relations.

Thus, if sentences containing proper names have sense properties and relations, the theory of sense to which neo-Millians are committed has to explain their sense properties and relations. We have already noted that some sentences containing proper names have the sense property of being meaningful. (5c) is an example. We have also noted that some sentences containing proper names are nonsynonymous. (5a) and (5b) are examples. Again, we have noted that some sentences containing proper names are analytic and some analytically entail others. (4) is an example of analyticity, while (2) and (3) provide an example of analytic entailment. Further, (21) and (22) are examples of sentences that are ambiguous (between an analytic and a synthetic sense—see footnote 11).

(21) Boys are bearers of 'Boy'.
(22) Smiths are bearers of 'Smith'.

Finally, (23) is synthetic and (24) contradictory.

(23) John Locke is a bearer of 'Giovanni Locke'.
(24) John Stuart Mill is not a bearer of 'John Stuart Mill'.

Compositionality enables us to drive the argument down from sentences containing proper names to the proper names they contain. An explanation of how it is that a sentence has the sense properties and relations it has is an explanation of how it is that the sentence has the sense(s) it has. For example, an explanation of a sentence being meaningful is an explanation of its having at least one sense, and an explanation of a sentence being ambiguous is an explanation of its having two or more senses. Now, since meaning is compositional, the explanation of a sentence's having the sense(s) it has (idioms notwithstanding) is an explanation of how those sense(s) are formed compositionality from the senses of its lexical items and its syntax. Thus, an explanation of the meaningfulness or ambiguity of a sentence is an account of how its sense(s) are formed compositionally from the senses of its lexical items.

If the proper name 'Santa Claus' in (5c) had no sense to contribute to the compositional meaning of the sentence, (5c) would not have a sense and would not be meaningful. If the proper names 'Boy' and 'Smith' in (21) and (22) had no sense to contribute to the compositional meaning of those sentences, these sentences would not have two senses and would not be ambiguous. Hence, the fact that (5c) has a sense and that (21) and (22) each have two senses means that their subjects, the proper names 'Santa Claus', 'Boy', and 'Smith', have a sense. The story is the same in the case of the explanation of the sense properties and relations of (2), (3), (4), (23), and (24). Since compositionality commits a theory of sense to ascribing senses to proper names in such sentences in order to account for their sense properties and relations, those who are committed to a theory of sense are committed to the descriptivist thesis that proper names are meaningful.[57]

Thus, in acknowledging the meaningfulness of common nouns and other expressions including improper names, neo-Millians commit themselves to a theory of sense that explains sense properties and relations of sentences. The fact that sentences containing proper names have sense properties and relations means that the theory has to explain such sentences, and, given the compositional nature of explanation, its explanations must ascribe senses to the proper names in sentences. Given compositionality, neo-Millians are committed to a theory of sense that says proper names have a sense. Hence, neo-Millians hold a philosophical theory that claims proper names have no sense and a linguistic theory that claims the opposite.

---

[57] The senses that a theory of sense commits us to saying that proper names have must be instances of (1')(i). Nothing less accounts for the contribution of a proper name to the compositional explanation of the sense properties and relations of sentences. Ascribing the sense 'the thing which is a bearer of "Socrates"' to the proper name "Socrates" explains why (4) is trifling, why its predicate merely repeats what its subject already says. Ascribing a sense which is an instance of (1')(i) to 'John Locke' explains the nonanalyticity (23), since, on this ascription, the sense of (23) has the form 'a bearer of '*N*' is a bearer '*M*''. And, ascribing a sense which is an instance of (1')(i) to 'John Stuart Mill' in (24) explains the contradictoriness of (24), since, on this ascription, the sense of (24) has the form 'a bearer of '*N*' is not a bearer of '*N*''. Correspondingly, nothing more than an instance of (1')(i) would satisfactorily represent the contribution of a proper name to the compositional meaning of a sentence, since nothing more is needed to account for its sense properties and relations. Accordingly, simplicity considerations limit ascriptions to instances of (1')(i).

V EPILOGUE

Referring to some of the major arguments that proper names have a sense, including Frege's argument that sentences like (5a) are analytic whereas sentences like (5b) are not, Kripke says

> It would be nice to answer all of these arguments I am not entirely able to see my way clear through every problem of this sort that can be raised Nevertheless, I think it's pretty certain that the view of Frege and Russell is false [58]

Kripke was right to think that classical descriptivism is false, but wrong to think it possible "to answer all of these arguments " Descriptivism being true, they are largely persuasive arguments for a true conclusion In attempting to defend Millianism with Millian arguments, Kripke was trying to do the impossible He was trying to refute the descriptivist thesis that proper names have a sense with arguments that work only against the Fregean thesis that they have a reference-determining sense

As we have seen, Kripke's faith that Millian arguments work against the descriptivist thesis rests on his equation of sense with Fregean sense and intensionalism with Fregean intensionalism Kripke is by no means alone in equating them Wittgenstein, Quine, and most contemporary analytic philosophers also equate them Being widespread, the equation has had widespread consequences Elsewhere, I have argued that it has limited the approaches that philosophers have considered in many controversies besides the controversy about names, and, further, that, in those controversies, too, the failure to consider non-Fregean intensionalism has prevented an intuitively satisfying resolution [59] If intuitively satisfying resolutions of those other controversies also come from introducing non-Fregean intensionalism into them, then the foundations of analytic philosophy—the tradition which comes down to us from Frege, Russell, Wittgenstein, and Quine and within which most Anglo-American philosophers now work—is due for a much-needed re-examination

JERROLD J KATZ

Graduate Center/City University of New York

---

[58] *Naming and Necessity*, p 29

[59] *The Metaphysics of Meaning* (Cambridge MIT, 1990), "Analyticity, Necessity, and the Epistemology of Semantics", "The Problem in Twentieth-Century Philosophy," this JOURNAL, XCV, 11 (November 1998) 547-75, *Sense, Reference, and Philosophy*

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO 4, APRIL 2001

28|5|2001

SYSTEMATICITY AND THE COGNITION OF STRUCTURED DOMAINS*

The current debate over systematicity concerns the formal conditions a scheme of mental representation must satisfy in order to explain the systematicity of thought[1] The systematicity of thought is assumed to be a pervasive property of minds, and

---

* Versions of this article were presented before the Society for Philosophy and Psychology (June 1999), CogSci99 (July 1999), and the American Philosophical Association (April 2000) and at various universities We thank audiences and commentators for their helpful remarks Research was funded by National Science Foundation Grant # 9976739

[1] The debate was instigated by Jerry A. Fodor and Zenon W Pylyshyn, "Connectionism and Cognitive Architecture," *Cognition*, xxviii (1988) 3-71 In the present article, we mainly focus on that paper and on Fodor and Brian P McLaughlin, "Connectionism and the Problem of Systematicity Why Smolensky's Solution Doesn't Work," *Cognition*, xxxv (1990) 183-205, and Paul Smolensky, Géraldine Legendre, and Yoshiro Miyata, "Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition," (Tech Report 92-08, Institute of Cognitive Science, University of Colorado) See footnote 2 for other papers detailing Fodor's position and footnote 5 for papers detailing Smolensky's Other important participants in the debate were Kenneth Aizawa, "Explaining Systematicity," *Mind and Language*, xii (1997) 115-36, Keith L Butler, "Towards a Connectionist Cognitive Architecture," *Mind and Language*, vi (1991) 252-72, Butler, "Compositionality in Cognitive Models The Real Issue," *Philosophical Studies*, LXXVIII (1995) 153-62, David J Chalmers, "Connectionism and Compositionality Why Fodor and Pylyshyn Were Wrong," *Philosophical Psychology*, vi (1993) 305-19, Nick Chater and Mike Oaksford, "Autonomy, Implementation and Cognitive Architecture A Reply to Fodor and Pylyshyn," *Cognition*, xxiv (1990) 93-107, Robert Cummins, "Systematicity," this JOURNAL, XCIII, 12 (December 1996) 591-614, Robert F Hadley, "Systematicity in Connectionist Language Learning," *Mind and Language*, ix (1994) 247-72, Hadley, "Systematicity Revisited," *Mind and Language*, ix (1994) 431-44, Hadley and Michael B Hayward, "Strong Semantic Systematicity from Hebbian Connectionist Learning," *Minds and Machines*, vii (1997) 1-55, Hadley, "Cognition, Systematicity, and Nomic Necessity," *Mind and Language*, xii (1997) 137-53, Terence Horgan and John Tienson, "Structured Representations in Connectionist Systems?" in Steven Davis, ed , *Connectionism Theory and Practice* (New York Oxford,

0022-362X/01/9804/167–85

can be characterized (roughly) as follows anyone who can think *T* can think systematic variants of *T*, where the systematic variants of *T* are found by permuting *T*'s constituents So, for example, it is an alleged fact that anyone who can think the thought *that John loves Mary* can think the thought *that Mary loves John*, where the latter thought is a systematic variant of the former

I BACKGROUND

The systematicity of thought itself, of course, cannot be directly observed But it is easy to see why it is widely assumed Anyone who can *understand* a sentence *S* can understand its systematic variants Since understanding *S* requires having the thought it expresses, anyone who can think the thought expressed by a sentence *S* can have the thought expressed by the systematic variants of *S* The systematicity of all thought expressible in language seems to follow if you can understand 'John loves Mary', you must be capable of the thought *that John loves Mary* Since anyone who can understand 'John loves Mary' can understand 'Mary loves John', it follows that anyone who can have the thought *that John loves Mary* can have the thought *that Mary loves John*

To derive the thought one must have to understand the systematic variants of 'John loves Mary', permute the words in 'John loves Mary' (while maintaining grammaticality), and preface the result with 'that' This leads to what we might call the *orthodox position* which Jerry A Fodor, Zenon W Pylyshyn, and Brian P McLaughlin[2] defend concerning the explanation of systematicity, namely, that it is best understood as involving two parts

---

1991), pp 195-228, Robert J Matthews, "Three-concept Monte Explanation, Implementation, and Systematicity," *Synthese*, CI (1994) 347-63, Matthews, "Can Connectionists Explain Systematicity?" *Mind and Language*, XII (1997) 154-77, Lars F Niklasson and Tim van Gelder, "On Being Systematically Connectionist," *Mind and Language*, IX (1994) 288-302, van Gelder, "Compositionality A Connectionist Variation on a Classical Theme," *Cognitive Science*, XIV (1990) 355-84

[2] Fodor, Pylyshyn, and McLaughlin's position is fully laid out in the following articles Fodor and Pylyshyn, "Connectionism and Cognitive Architecture", Fodor and McLaughlin, "Connectionism and the Problem of Systematicity Why Smolensky's Solution Doesn't Work", Fodor, "Connectionism and the Problem of Systematicity (Continued) Why Smolensky's Solution Still Doesn't Work," *Cognition*, LXII (1997) 109-19, McLaughlin, "Systematicity, Conceptual Truth, and Evolution," in Christopher Hookway and Donald Peterson, eds , *Philosophy and the Cognitive Sciences*, Royal Institute of Philosophy, Supplement no 34 (New York Cambridge, 1993), pp 217-34, McLaughlin, "The Connectionism/Classicism Battle to Win Souls," *Philosophical Studies*, LXXI (1993) 163-90 Hereafter, 'FPM' is used in the text to refer to the authors' position, result, and the like, otherwise, their full names are used

(1) *The representational theory of thought* having the thought that *p* is having a *p*-expressing mental representation in a certain cognitive role

For example, having a belief *that p* amounts to having a mental representation *that p* in the belief box[3]

(2) Mental representation is "classical" mental representation has a language-like combinatorial syntax and associated semantics[4]

Putting these two parts together, we get that anyone who can think *that John loves Mary* can think *that Mary loves John,* since (i) thinking 'Mary loves John' involves tokening a representation of the proposition *that Mary loves John,* and (ii) that representation has constituents corresponding to Mary, John, and the relation of loving, which can simply be permuted to yield a representation, and hence a thought, corresponding to the proposition *that John loves Mary* Fodor, Pylyshyn, and McLaughlin thus conclude that the human system of mental representation must be "classical," that is, a language-like scheme having the familiar kind of combinatorial syntax and associated semantics first introduced by Alfred Tarski (*op cit*)

An unfortunate consequence of the way Fodor, Pylyshyn, and McLaughlin have characterized the systematicity of thought is that any theory that accounts for understanding every sentence will account for systematicity trivially if one can understand every sentence, one can understand every systematic variant of any given sentence If the domain of sentences to be understood is finite, nonclassical schemes could be complete in the relevant sense, and hence account for systematicity For instance, a look-up table that uses arbitrary names to represent each sentence of the domain would account for it

A natural reply to this point would be to claim that only representational schemes employing something like classical combinatorics could be complete for an unbounded domain like the domain of thought-expressing sentences in a natural language But Paul Smolensky, Géraldine Legendre, and Yoshiro Miyata[5] have proven that for every classical parser—that is, a parser defined over classical representations—there exists a tensor-product network that is weakly (input-output) equivalent to the classical parser but does not employ classical representations.[6] Thus, it appears that the classical explanation and the tensor-product explanation explain the systematicity of thought equally well.

---

[3] This assumption of the representational theory of thought is controversial One might think that a person can have the thought that *p* without having the representation that *p* See Cummins, *Representations, Targets, and Attitudes* (Cambridge MIT, 1996)

[4] We call these schemes "classical" in what follows (one could also call them "Tarskian") See Alfred Tarski, "The Concept of Truth in Formalized Languages" (1936), in his *Logic, Semantics, Metamathematics* (New York Oxford, 1956), pp 152-278, "The Semantic Conception of Truth" (1944), in Herbert Feigl and Wilfrid Sellars, eds , *Readings in Philosophical Analysis* (New York Appleton, 1949), pp 52-84

There are two *philosophical* responses to the SLM result that we want to surface briefly and put aside. The first, due to Georg Schwarz,[7] is that networks like that designed by Smolensky, Legendre, and Miyata do not really have an unbounded competence. The second, suggested by Fodor, Pylyshyn, and McLaughlin, and discussed briefly by Cummins,[8] is that, while both parsers[9] can account for the systematicity data, the SLM explanation is unprincipled because the tensor-product parser is handcrafted to achieve systematicity, whereas the classical parser is not. We discuss these briefly in turn.

---

[5] Smolensky, Legendre, and Miyata's position is fully laid out in the following articles: Smolensky, "The Constituent Structure of Connectionist Mental States," *Southern Journal of Philosophy Supplement*, XXVI (1987): 137-60; Smolensky, "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," *Artificial Intelligence*, XLVI (1990): 159-216; Smolensky, "Connectionism, Constituency and the Language of Thought," in Barry Loewer and Georges Rey, eds., *Meaning in Mind: Fodor and His Critics* (Cambridge: Blackwell, 1991), pp. 201-27; Smolensky, "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture," in Cynthia MacDonald and Graham Macdonald, eds., *Connectionism: Debates on Psychological Explanation* (Cambridge: Blackwell, 1995); Smolensky, Legendre, and Miyata, "Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition." Hereafter, 'SLM' is used in the text to refer to the authors' position, result, and the like; otherwise, their full names are used.

[6] Tensor-product encoding is a general-purpose technique for binding fillers to roles (object to properties/relations). An activation vector representing the binding of a filler *f* to a role *R* is obtained by taking the tensor product of the vectors representing *R* and *f*. The resulting vector can then be added to (superimposed on) others representing bindings of fillers to roles, yielding a single vector that represents the binding of many fillers to many roles. Smolensky, Legendre, and Miyata use this technique to construct recursive representations of binary trees. Matrices effecting the vector operation for constructing a tree from its left child *p* and right child *q* (or for decomposing it) are then defined, making possible direct implementation in one layer of connection weights.

This connectionist representation of trees enables massively parallel processing. Whereas in the traditional sequential implementation of LISP, symbol processing consists of a long sequence of car, cdr, and cons operations, here we can compose together the corresponding sequence of W_{car}, W_{cdr}, W_{cons0}, W_{cons1} operations into a single matrix operation. Adding some minimal nonlinearity allows us to compose more complex operations incorporating the equivalent of conditional branching (Smolensky, Legendre, and Miyata, *ibid.*). The matrices can then be straightforwardly combined to yield a single layer of connection weights that implement parse-dependent predicates and functions.

[7] "Connectionism, Processing, Memory," *Connection Science*, I (1992): 207-26.

[8] "Systematicity."

[9] Smolensky, Legendre, and Miyata do not in fact construct a parser. What they do is show how to construct networks effecting parse-dependent functions and

Unbounded competences in connectionist networks. Schwarz has argued that connectionist networks that are not simply implementations of classical architectures cannot have unbounded competences. According to Schwarz, we take an ordinary calculator to have an unbounded competence because it employs perfectly general numerical algorithms for addition, multiplication, subtraction, and division. Consequently, its competence is limited only by time and memory.[10] We can add more memory to the calculator, and let it run longer, without altering the algorithms it exploits. But, Schwarz argues, the same does not hold of a connectionist calculator, for the only way to add memory is to add more nodes, or to add precision to the nodes. Doing either of these will require retraining the network, however, which amounts to programing a new system that, in effect, executes a different algorithm. The fact that processing and memory are fully integrated in connectionist networks blocks the standard idealization away from memory limitations that licenses attribution of unbounded competences to classical systems whose controlling algorithms remain unchanged with the addition or subtraction of memory.

We believe this argument should be rejected for reasons given by Cummins.[11] The culprit is the assumption that the function computed is a function from input to output rather than a function from input and initial state to output and final state. The input-to-output relation exhibited by a classical parser, for example, is a function of its stored knowledge. Change what the system knows and the same input will yield a different output. Therefore, the input-output relation is seldom a function. This is particularly obvious when we consider learning algorithms. These cannot be conceived as functions pairing inputs and outputs, for the whole point of a learning algorithm is to replace ineffective input-output pairings with better ones. By definition, learning algorithms remain constant over changes in input-to-output relations.

---

predicates (for example, active-passive) that correspond precisely to arbitrarily complex parse-dependent LISP functions/predicates. In the context of the current discussion, this amounts to showing that classical representation is not required for systematic and productive capacities defined over the constituent structure of a sentence.

[10] It is also limited, ultimately, by wear and tear, but this raises no special problem for connectionists.

[11] "Systematicity."

We accept the principle that a different function computed implies a different algorithm executed But the function computed by a connectionist system must be conceived as a function from an activation vector and point in weight space to another activation vector-weight space pair From this point of view, we do not build a new network when we change weights any more than we build a new rule-based system when we change its stored knowledge

Connectionist data coverage is "unprincipled" The idea here is that classical representational schemes *predict* systematicity, whereas connectionist schemes at best *accommodate* it

To get a concrete sense of this objection, suppose a classical system generates a phrase marker for 'John loves Mary' Since 'Mary loves John' has precisely the same phrase marker, except that 'John' and 'Mary' have exchanged positions, the system is bound to be able to parse 'Mary loves John' if it can parse 'John loves Mary' the grammatical structure is the same, and so are the lexical resources

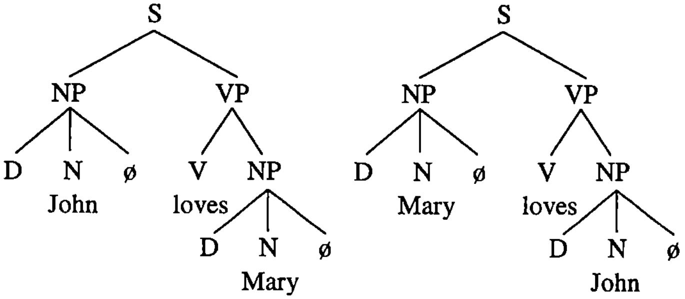


Figure 1

By contrast, so the argument goes, a connectionist network could be trained to parse 'John loves Mary' but not 'Mary loves John'

An obvious reply is that while a classical system might be programed to exhibit the very same incapacity, the resulting system would be obviously ad hoc The incapacity does not appear ad hoc in the connectionist framework, however, since the representations of the sentences do not themselves have constituents corresponding to the relevant lexical items It might seem, therefore, that there is no reason why the network would have to process the two representations in a similar way They are simply different activation vectors, and hence there will be no special problem with a training set that simply declares the second parse unacceptable Bias against a systematic variant of an acceptable sentence will present no learning problem different in principle from bias against any other unacceptable sentence—for example, 'Loves Mary John' A classical system, however, will give both parses if it gives either, unless some special rule is added to the grammar to block one or the other There will be no rule or process in the deviant connectionist system that is special to this case The same setting of weights that deals with every other sentence, whether acceptable or unacceptable, will deal with the deviation There will, in short, be no principled distinction between the way it rejects 'Mary loves John' and the way it rejects 'Loves John Mary'

Here is a different but related way to see the alleged problem Imagine a classical system that can parse 'John loves Mary' but cannot parse 'John despises Mary' because 'despises' is not in its lexicon It seems intuitively plausible to suppose that a simple learning algorithm will add 'despises' to the lexicon, and that this will suffice to yield the new parse The SLM result shows that for each of these classical parsers, there is a correlate, weakly equivalent, connectionist parser It is silent, however, about the relation between these two correlate connectionist parsers In particular, it is silent about whether the correlate connectionist parser that handles 'despises' is in any sense a natural extension of the one that does not For all the SLM result shows, the connectionist parser that handles 'despises' would have to be built from scratch and not a simple modification of the one that does not By itself, the SLM result does not show that the networks corresponding to the pre- and post-learning stages lie on any common connectionist learning trajectory It shows only that *any* set of parses that can be captured by a classical system can be captured by a connectionist network This at least suggests the possibility that the connectionist network does not parse the systematic variants of *S because they are systematic variants of an acceptable sentence,* but rather that it parses them simply because they are among the sentences in the target set defined by the classical system to be emulated We have, in short, no reason to think that all the systematic variants of *S* are parsed in the same way Rather, the prediction that the network will parse a systematic variant of *S* is derived simply from the facts (i) that the systematic variants of *S* are acceptable if *S* is, and (ii) that the network parses all the acceptable sentences [12]

---

[12] This problem is an instance of a more general problem we address elsewhere and will be pursued in further publications See Pierre Poirier, Cummins, James Blackmon, David Byrd, Martin Roth, and Schwarz, "The Epistemology of Nonsymbolic Cognition Atomistic Learning and Forgetting" (Tech Report Phil99-3, University of California/Davis, 1999), see also Michael McCloskey and Neil J Cohen, "Catastrophic Interference in Connectionist Networks The Sequential Learning Problem," *The Psychology of Learning and Motivation,* XXIV (1989) 109-65

We shall come back in a later section to the merits (and demerits) of the objection that connectionist data coverage is unprincipled. But notice that, at this point, the debate has shifted from empirical considerations of what best covers the data to philosophical considerations of what constitutes a principled explanation. We propose, therefore, to reformulate the issue in the hope of finding a way of retaining its essentially empirical cast.

II. THE ISSUE REFORMULATED

The fact that a system cognizes a domain will manifest itself in a variety of psychological effects. By an *effect*, we mean a nomic regularity in behavior (what Ruth Garrett Millikan[13] calls a law *in situ*, that is, a law that holds of a system in virtue of the special structure and organization of that system). For instance, the fact that humans cognize the color domain is manifested by the fact that humans make such and such discriminations (and fail to make others), by the fact that they will eat some foods and not others, and so on.

We take it as uncontroversial that some domains are cognized by grasping their underlying structure. For example, we can recognize the melody of "Mary Had a Little Lamb" whether it is played by an orchestra or on a kazoo, regardless of what key it is played in, and, within limits, regardless of tempo. Sensitivity to melody across differences in timbre, tempo, and key suggests that we process a melody by processing information about the structure and arrangement of notes in a composition. Word problem-solving in algebra is another case where grasping the underlying structure of a domain is required for cognizing the domain. Students are better able to solve word problems involving distances, rates, and times when they see the problems as having the same structure as problems with which they are already familiar. For example, students who can solve problems involving wind speeds but have difficulty solving problems involving current speeds are able to perform well on the latter when it is

---

Connectionist systems using distributed representation cannot learn atomically, that is, they cannot add single representations, such as the predicate 'despise', to their belief box or knowledge base without having to relearn everything they previously learned. Surely, this is not the way we acquire new representations in many domains. (But, for a possible solution, see James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly, "Why There Are Complementary Learning Systems in the Hippocampus and the Neocortex: Insights from the Success and Failures of Connectionist Models of Learning and Memory," *Psychological Review*, cii (1995): 419-57.)

[13] *Language, Thought, and Other Biological Categories* (Cambridge: MIT, 1984).

pointed out that current is analogous to wind [14] Only when students recognize an underlying structure with which they are familiar do they seem able to give the correct answer (this suggests that sensitivity to structure is learned in some cases, or at least affected by prior knowledge) We also take it to be uncontroversial that some domains are cognized without grasping any significant underlying structure knowing the capitals of twenty states in the United States is no help in divining the capital of an unstudied state Our ability to learn the state capitals does not depend on being sensitive to underlying structural features of states and their capitals With respect to the previous distinction, we call an effect a *systematicity effect* if it is a psychological effect that implies sensitivity to the structure of the domain cognized [15]

Is it possible to draw any conclusions about the form of mental representation from the presence of systematicity effects? We think it is Recall that our main methodological objective here is to keep the issue at an empirical level in the hope that it can be solved there Our first conclusion proceeds from the observation that not all systematicity effects are created equal some may be computed by the system, and some may be incidental effects of whatever algorithm is being computed We call the first type of systematicity effect *primary* and the second *incidental*, and in the next section we address the distinction and what it tells us about mental representation [16] Our second conclusion rests on a finer analysis of the SLM tensor-product representations Unlike classical representations that preserve information about the structure of represented elements by actually sharing that structure, tensor-product representations do not share structure with what they represent yet still manage to preserve structural information and make it available to processors This allows a possible solution to a problem raised by Cummins[17] against classical representations—namely, that they cannot possibly share structure with every domain in which we find systematicity effects—but the price for that solution may be one Fodor, Pylyshyn, and McLaughlin are not ready to pay We address this issue in section IV

---

[14] Denise Cummins, "Role of Analogical Reasoning in the Induction of Problem Categories," *Journal of Experimental Psychology Learning, Memory, and Cognition*, v (1992) 1103-24

[15] These examples are meant to represent the extreme cases There are likely to be cognitive tasks that require grasping varying degrees of structure depending on the domain, so the distinction between structured and unstructured is not all or nothing Also, some cognitive tasks will be hybrid cases, where some aspects require grasping structural information while others do not Cognizing language appears to be an example, since mastery of the primitive lexicon is surely more like learning the state capitals than like mastering the syntax

One must also be careful not to confuse structure in the domain with structure in the way a problem or question is posed When a teacher requires that I learn the capital of California, the structure of my answer is determined in part by the structure of the question itself Similarly, one can imagine a state capital learning device that represents states as one place predicates such that only one capital could be matched with any state In this case the structure of the answer is partly determined by the way states are represented in the system

[16] The distinction between primary and incidental systematicity effects is introduced in Cummins, "Systematicity," and also discussed in Cummins, "'How Does It Work?' versus 'What Are the Laws?' Two Conceptions of Psychological Explanation," in Frank C Keil and Robert A Wilson, eds , *Explanation and Cognition* (Cambridge MIT, 2000), pp 117-44

III PRIMARY VERSUS INCIDENTAL SYSTEMATICITY EFFECTS

The distinction between primary and incidental effects captures the important fact that a system's behavior results not only from the function it computes, but from a variety of other factors as well Compare a system that multiplies by partial products with a system that computes products through successive addition What the two systems have in common is the *multiplication effect,* that is, the fact that they both produce the same products from the same arguments They differ, however, in how long it takes them to compute the value for a given argument In the case of the successive adder, the response time is roughly proportional to the size of the multiplier computing ten times some number will take approximately twice as long as computing five times that number The other system, in contrast, does not display such a linearity effect Its response time is, roughly, a step function of the number of digits in the multiplier, indicating how many partial products need to be added in the end

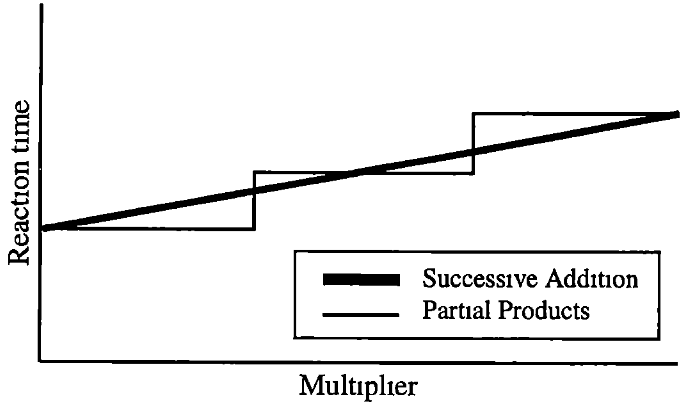
The two systems have the same primary systematicity effects, since every argument pair leads to the same value in both systems they are both multipliers But their incidental systematicity effects are different For instance, the partial product multiplier will take approximately the same amount of time to multiply N by 10 as it will to multiply N by 99, whereas the successive addition multiplier will take roughly 10 times longer to multiply N by 99 as it will to multiply N by 10 In general, two systems that compute the same function using different algorithms will display different incidental effects, although they will be weakly equivalent in Pylyshyn's sense[18]

One important source of incidental effects thus lies in the algorithm producing the primary effects But systems that implement the *same* algorithm may exhibit different incidental effects If the under-

---

[17] "Systematicity "

[18] *Computation and Cognition* (Cambridge MIT, 1984)



Products are computed (primary)
Reaction time is incidental

Figure 2

lying hardware is sufficiently different, they may operate at greatly different speeds, as is familiar to anyone who has run the same program on an Intel 486 running at 50 MHz and on a Pentium III running at 750 MHz A significant difference in response time between two systems, in other words, does not entail that they implement different algorithms, it may be a direct result of a difference at the implementation level Hence, two systems that have identical primary effects may exhibit different incidental effects either because they compute different algorithms or because the identical algorithms they compute are implemented in different hardware

Incidental effects are not restricted to the complexity profiles of the algorithm or the constraints imposed by the implementing matter on the execution of the algorithm Continuous operation will heat up the calculator, an incidental effect that is normally irrelevant to the system's performance Nonetheless, when things get too hot, the system will start malfunctioning and ultimately break down As a consequence, two systems that differ in performance do not necessarily compute different functions, one of the systems may simply have been subject to extraneous factors, such as overheating, fatigue, or attention deficit disorder By the same token, the general physical makeup of a system also determines the extent to which environmental conditions will have an impact on the system's ability to operate The occurrence of a strong magnetic field, for example, will interfere with the operation of an electronic calculator but not with that of an abacus In the case of single-purpose machines, this last kind of incidental effect is primarily relevant for explaining why a system fails to display its primary effects on certain occasions But in the case of more complex systems, such effects have proven useful as an explanatory tool as well Consider the fact that cognitive processing in the human brain is generally correlated with increased metabolic activity, a fact that has proved critical for the use of imaging technology (for example, PET, fMRI) that studies which cortical areas are involved in the processing of these tasks Yet metabolic activity as such is not specific to the neural implementation of whatever computations are performed

It is evident that the effects of implementation details and environment complicate inferences from sameness of incidental effects to sameness of underlying functional architecture In what follows, however, we shall make use only of the inference from differences in incidental effects to differences in functional architecture in cases in which implementational and environmental influences are not at issue

Now, Smolensky, Legendre, and Miyata have proven that a connectionist tensor-product parser and the classical parser can exhibit the same primary effects (they are weakly equivalent, like our two multiplication algorithms) Any parse a classical parser computes can be computed by a corresponding connectionist parser If systematicity effects are primary effects, then Smolensky, Legendre, and Miyata have demonstrated, mathematically, that systematicity effects in language parsing can be accounted for without any appeal to classical representations Hence, if the systematicity effects at issue are primary effects, then nothing can be concluded concerning whether the mental representations involved are classical or connectionist in form As we saw, the consequence of that empirical deadlock has been to turn away from empirical considerations to philosophical issues concerning what constitutes a principled explanation The SLM tensor-product parser employs a different algorithm than the classical parser, however, and thus the two systems are bound to exhibit different incidental effects If the systematicity effects observed are incidental, then, barring cosmic coincidence, at most one of the two parsers can account for them In particular, the classical explanation can gain leverage over the SLM explanation if the form in which information is represented is important to explaining the effect An incidental effect of using classical representations in language processing might, for instance, involve priming Since the lexical items and sentential structure of a recently processed sentence *S* may be easily accessible (they may still be in short-term memory, for instance), the recent processing of *S* may allow the system to process systematic variants of *S* faster than it would process sentences that are not systematic variants of *S* An incidental effect of using connectionist representations is the flat temporal profile of parsing with a tensor-product parser Since the latter will process all sentences, no matter how complex, in one step, it will process all of them in exactly the same amount of time Should the empirical evidence show that subjects exhibit the kind of priming effect described above or a temporal profile different from the one we would expect from the use of a tensor-product parser, then, weak equivalence notwithstanding, Fodor, Pylyshyn, and McLaughlin have a good argument in favor of their claim that classical representations better explain the systematicity of thought Of course, should the evidence show that the incidental effects are those one would expect from an SLM-type parser, then it would seem that classical representations are not the source of the systematicity of thought

Fodor, Pylyshyn, and McLaughlin and their supporters cite no evidence from incidental systematicity effects, nor do their opponents It seems likely that there are relevant effects reported in the literature on language processing, but they have not, to our knowledge, been brought to bear on this issue [19] It should be emphasized, however, that systematicity effects of the sort Fodor, Pylyshyn, and McLaughlin had in mind are clearly not incidental but primary[20] they do not argue from evidence that processing 'Mary loves John' makes it easier to process 'John loves Mary' to the conclusion that mental representation is classical They simply argue from the availability of a given thought (sentence understood) to the possibility of its systematic variants

IV REPRESENTATIONAL PLURALISM AND STRUCTURAL ENCODINGS

The inference from systematicity effects in language to classical representations involves three steps

(1) The observation of systematicity effects in language
(2) An inference from the presence of these effects to the conclusion that mental representations must preserve and carry information about the structure of the domain
(3) An inference from that conclusion to the further conclusion that

---

[19] We are in the process of searching the literature for such evidence Meanwhile, the fact that the SLM parser accomplishes all parses in a single step, regardless of the complexity of the sentence parsed, is surely suggestive

[20] This is put misleadingly in Cummins, "Systematicity," p 603 The point is that proponents of the systematicity argument are thinking of systematicity effects as primary when they should be thinking of them as incidental

> the information about structure must be carried by classical representations

Most will readily agree that systematicity effects in language can be observed We noted in the previous section that care should be taken to distinguish primary from incidental systematicity effects in language since only the latter will allow conclusions to be drawn about the nature of representations Step 2 is also uncontroversial In unbounded domains (see above), how else can these effects be produced? It is step 3 of the inference, from the preservation of structural information to the necessary presence of classical representations, that we wish to address here

Sensitivity to the structure of a domain is best explained by a scheme of mental representation that carries information about the structure of that domain One such scheme involves structural representations of the domain in question A representation *R* is a *structural representation* of its target just in case *R* and its target share structure Scale models, photographs, and maps are typical examples of structural representations As Cummins[21] points out, it is natural to explain the systematicity effects we find in language, vision, and audition by positing structural representations of each of the respective domains The obvious virtue of structural representation is that structural transformations and permutations of the representations yield representations of the systematic variants of the corresponding targets in the cognized domain For every operation defined over elements in a domain there can be a corresponding operation defined over elements in the representational scheme Thus, the representation of an item in the domain can be used to construct a representation of its systematic variants The sensitivity to structure that is required to cognize certain domains is accomplished by actually having the structure of the domain in the representations We are now in a position to appreciate that the FPM classical representations are a case of a scheme of structural representation for the linguistic domain The idea is that, in order to be sensitive to the combinatorial syntax and associated semantics of a language, there must be a system of internal representations that has the same (or corresponding) syntactic and semantic features

This way of explaining systematicity effects, however, limits classical schemes to the explanation of systematicity effects exhibited by the cognition of language-like domains, since these are the only domains that classical schemes can structurally represent This leaves the systematicity effects we find in other differently structured domains to be explained by appeal to nonclassical schemes Since we apparently grasp the underlying structure of domains structurally distinct from language, we would require structurally distinct representational schemes to cope with each of them in the same way that classical schemes are supposed to facilitate coping with language Apparently, structural representation comes at the price of (perhaps massive) representational pluralism

---

[21] "Systematicity "

The problem of representational pluralism was first noticed by Cummins (*ibid*) Roughly, representational pluralism is the idea that for every differently structured domain for which we show a systematicity effect, we employ a scheme of mental representation that shares structure with that domain Thus, if we exhibit systematicity effects in cognizing three differently structured domains, we employ at least three differently structured schemes of mental representation Fodor, Pylyshyn, and McLaughlin can make a plausible argument from systematicity effects in language processing to the conclusion that some mental representation is classical There are other systematicity effects, however, such as those found in vision and audition, that cannot be accounted for by structural representations if all mental representation is classical, since the structure of the domains in language, audition, and vision are different Since the inference from systematicity effects in domains other than language to nonclassical, structural representations is on par with the inference from systematicity effects in language to classical representations, the systematicities in vision and audition are good evidence for some nonclassical, structural representations Thus, if the FPM inference is sound, it constitutes a good argument for representational pluralism The only way Fodor, Pylyshyn, and McLaughlin can retain the idea that mental representation is monistic is by allowing that some systematicity effects can be adequately explained without recourse to a corresponding scheme of structural representations Some systematicity will have to be explained by appeal to mere *encodings* An encoding of a domain *D* is a mapping of the members of *D* onto the representations in a scheme *R* whose members do not share the structure of their images in *D* Classical representations structurally represent linguistic structure, but they only encode the structure of music By allowing some encoding, however, friends of Fodor, Pylyshyn, and McLaughlin would forfeit the objection that connectionist data coverage is unprincipled, since encoding, on their view, forces us to be unprincipled somewhere So it seems that either the objection must go, or Fodor, Pylyshyn, and McLaughlin are forced to accept representational pluralism

The SLM parser shows, however, that representational monism is compatible with an adequate account of primary systematicity effects

To see this clearly, we need a taxonomy of representational schemes that (i) makes evident what property holds of structural representations in virtue of which they account for primary systematicity effects, and (ii) identifies an alternative sort of scheme that can be seen to have this critical property

Our taxonomy relies on two distinctions The first distinction has to do with whether a representational scheme represents its domain by making available representations that share structure with the items of the domain represented Shared structure involves two things First, the representations have constituents that represent the constituents of the domain Second, these representational constituents are structurally related in a way that represents the ways in which the constituents of the content are structurally related We call any scheme in which representations represent items in its target domain in virtue of sharing structure with the things they represent a *structural representational scheme* for that domain We call schemes that represent a domain, but not in virtue of shared structure, an *encoding* of the domain So the first distinction separates entities that represent in virtue of shared structure, that is, structural representations, from those which represent by some other means, that is, encodings

The second distinction we need requires the notion of a *recovery function* A recovery function is a function from a representational scheme to a domain Intuitively, these functions allow one to recover contents from representational schemes Recovery functions are not necessarily employed by cognitive systems, they may remain purely theoretical constructs with which we may interpret representational schemes We distinguish between recovery functions which are *systematic* and those which are not A recovery function is systematic if there is a general/productive algorithm for mapping the scheme to its contents In such a case, the function does not need to be defined by a list of pairs each containing an element of the scheme and a content If the recovery function is systematic, then the representational scheme is structural If the recovery function is not systematic, then the representational scheme is pure or arbitrary

	Systematic Recovery Function	No Systematic Recovery Function
Do not share structure	Structural Encodings	Pure Encodings
Share structure	Structural Representations	

This taxonomy allows us to distinguish three categories of representational entities structural representations, structural encodings, and pure encodings Schemes that belong to either of the first two categories are such that the information about the structure of the domain is systematically recoverable This is relevant because, if structure can be recovered systematically from a scheme, the scheme has preserved information about the structure of the domain The possibility of systematic recovery implies the preservation of structure

One obvious way of preserving structure is for the representational scheme simply to share the structure of the domain This is what structural representations do But this is not the only way Structural encodings preserve structure without sharing it Godel numbering and tensor-product schemes are both examples of structural encodings In the Godel scheme for encoding sentences, words are assigned natural numbers while their positions in the sentence are assigned prime numbers in ascending order A number *m* in position *n* yields the number n^m We can say that n^m stands for the word numbered *m* standing in the place numbered *n* This number is uniquely factorable into *n*, *m* times The Godel string of a sentence is determined by multiplying all these uniquely factorable numbers together This yields a number, expressed by a Godel string, which is uniquely factorable into a list of numbers An example of a Godel string is '1,093,500' which factors into $2^2 \times 3^7 \times 5^3$ which breaks down further into a list of 2s, 3s, and 5s The values of these numbers have been assigned to places, and the number of occurrences of any number has been assigned to words We may suppose the string encodes the sentence 'Mary loves John' where 'Mary', 'loves', and 'John' are assigned the numbers 2, 7, and 3, respectively The Godel string itself does not share structure with its content, however, this structure is systematically recoverable and thus the representation has preserved information about the structure of the sentence The SLM tensor-product representations have the same property The tensor-product activation patterns themselves do not share structure with their contents, but the existence of recoverability functions for such schemes entails that such schemes preserve structure

Structural encodings permit recovery but do not share structure with their domains Pure encodings do not permit systematic recovery and so do not preserve structure If, for example, we were to take the names of the state capitals as encodings of the names of the states, we would be employing a pure encoding scheme, for there is no systematic way by which the names of the states could be recovered from the names of the capitals[22] In this case, a look-up table is required, and no information about structure is preserved

Structural encodings, as opposed to pure encodings, are adequate to account for primary systematicity effects because, under such schemes, information about structure is preserved and that is all that is required This is, in fact, the property of structural encodings and structural representations in virtue of which it is possible to construct architectures using either one that exhibits systematicity effects To take our previous example, any system capable of processing 1,093,500 (the Godel string for 'Mary loves John') is also capable of processing 437,400 (the Godel string for 'John loves Mary') The parallel with the previous symmetric phrase markers is perfect The SLM tensor-product scheme is a structural encoding, not a pure encoding as Fodor, Pylyshyn, and McLaughlin seem to assume, and this is how Smolensky, Legendre, and Miyata are able to prove their otherwise puzzling result It is only when cognitive systems employ such structure-preserving schemes that they can be causally sensitive to the structure of the domain cognized and thus exhibit systematicity effects Systems employing pure encodings cannot exhibit such effects because there is no common form or structure to which the system could be causally sensitive

Once we see that structural encoding is adequate to account for systematicity effects—at least the primary ones—we are free to cleave to representational monism Since a tensor-product scheme like that employed by Smolensky, Legendre, and Miyata need not share structure with the domains it represents in order to account for primary systematicity effects, such a scheme could, in principle, account for primary systematicity effects in a variety of structurally distinct domains The same point can be made about classical representation since it can be used structurally to encode domains such as music that are structurally unlike language, a classical scheme could account for primary systematicity effects in nonlinguistic domains Since classical and nonclassical structural encodings are evidently on equal footing in this respect, we conclude again that there is no sound argument from primary systematicity effects in language to classical representation There is, of course, a sound argument from systematicity effects against pure encoding, but that should come as no surprise What is important to systematicity is preserving information about the relevant structure of the target domain It does not much matter how that information is encoded, provided an architecture exists that is capable of exploiting the information in that encoding Smolensky, Legendre, and Miyata demonstrate that connectionist networks can exploit such information as encoded in tensor-product activation vector schemes

---

[22] Another way to see the same point given a new capital name, there is no way to infer what state name it corresponds to

The staunch defender of classical representations may object at this point that there still is a methodological virtue in accepting classical representations-plus-pluralism over encodings-plus-monism, because it is the lawfulness of systematicity that needs to be explained and encodings just do not account for it But it not clear that they do not We agree that, with pure encodings, there could be minds that encode 'Mary loves John' without encoding 'John loves Mary', since the ability to encode the former does not imply the ability to encode the latter With structural encodings such as the SLM tensor-product scheme, however, matters are different Since the encodings are generated recursively from the filler and role vectors, the ability to encode 'Mary loves John' does imply the ability to encode 'John loves Mary', since the two encodings employ the same filler and role vectors The only difference between them is how the vectors get multiplied and added to create the respective tensor-product encodings of the two sentences

ROBERT CUMMINS, JAMES BLACKMON, DAVID BYRD,
PIERRE POIRIER, MARTIN ROTH, GEORG SCHWARZ
University of California/Davis

PERSONAL IDENTITY, MEMORY, AND CIRCULARITY:
AN ALTERNATIVE FOR Q-MEMORY*

I have many memories of my last summer's vacation, which I spent in the Rocky Mountains. For instance, I can recall vividly seeing a large Grizzly bear and her two cubs at a specific time and place (and distance). Since I was not used to wild animals larger than ducks, the event made a lasting impression on me.

Let us assume that, as a matter of objective fact, some person did see that bear at that precise time and place. Since I do not think I am deluded in remembering having seen that bear at that moment, that fact inclines me to believe that that person was me. This belief seems justified, but it can be justified for different reasons. It may be the case that, under normal circumstances, memory is reliable evidence for the identity of my present self and the bear-seeing person, however this identity is in fact constituted. But it may also be the case that memory itself coconstitutes identity over time, that is, that part of what it is or means for two time-indexed person-stages to be stages of one person is that the later stage has conscious access to the earlier stage via memory. In other words, memory can be either *indicative* or *constitutive* of personal identity over time.

The latter option is generally considered to be implied by a psychological criterion of personal identity over time. Contemporary philosophers[1] who defend such a criterion hold (1) that personal identity over time can be analyzed in terms of psychological continuity, and (2) that psychological continuity is coconstituted by memory. To be sure, memory is not thought to exhaust psychological continuity, but neither is it considered to be an optional ingredient. In fact, contemporary conceptions of psychological continuity evolved by adding various kinds of psychological connections between person-stages—connections such as those established by perpetuated

---

* Thanks to Marya Schechtman, Jan Bransen, Stefaan Cuypers, Joel Anderson, and Michael Smith for helpful comments. Work on it was sponsored by the Dutch Organization for Scientific Research (NWO).

[1] For example, Derek Parfit, *Reasons and Persons* (New York: Oxford, 1984), Sydney Shoemaker, "Persons and Their Pasts," in *American Philosophical Quarterly*, VII, 4 (1970): 269-85, Shoemaker (with R. G. Swinburne), *Personal Identity* (New York: Blackwell, 1984), Harold Noonan, *Personal Identity* (New York: Routledge, 1989), John Perry, "Personal Identity, Memory, and the Problem of Circularity," in Perry, ed., *Personal Identity* (Berkeley: California UP, 1975), pp. 135-55, David Lewis, "Survival and Identity," in Amelie Oksenberg Rorty, ed., *The Identities of Persons* (Berkeley: California UP, 1976), pp. 17-40.

0022-362X/01/9804/186–214

beliefs, values, and character traits or by relations between intentions and actions—to John Locke's[2] memory criterion of identity, this criterion is held to be too tight, but correct in spirit

Since memory is considered a (the) main ingredient of psychological continuity, one of the major objections against a psychological criterion of personal identity is the contention that memory presupposes personal identity This contention, the so-called *circularity objection,* is my topic here

I shall argue that a psychological criterion of personal identity can be rescued from the circularity objection only by accepting what has not even seemed to be an option to philosophers so far that memory is only indicative rather than constitutive of psychological continuity and hence, by a psychological criterion, of personal identity

In order to do this, I shall claim that the traditional solution to the objection, Sydney Shoemaker's Q-memory (section I), is seriously incomplete (section II) Next, I shall show that there is a kind of psychological continuity that is not (co-)constituted by memory and does not presuppose personal identity (section III) Memory, I shall argue, presupposes—and is hence indicative of—this kind of continuity (section IV) By analyzing personal identity in terms of this alternative kind of psychological continuity, the circularity objection can be avoided (section V) The new solution to the circularity objection I shall propose accepts and explains the intuition that memory presupposes personal identity—memory presupposes the alternative kind of psychological continuity, which, in turn, analyzes personal identity—without letting this stand in the way of a psychological criterion of personal identity

I THE CIRCULARITY OBJECTION AND Q-MEMORY

The circularity objection was originally formulated by Joseph Butler[3] to refute Locke's (purported) memory criterion of personal identity Although it is likely that Butler's point was merely to argue against the conflation of epistemic and metaphysical criteria of identity,[4] the objection has been interpreted in far more sophisticated ways since then [5] One way of presenting the sophisticated version of the objection is by means of the distinction between *veridical* and *nonveridical* memory

---

[2] *An Essay Concerning Human Understanding* [1670], A D Woozley, ed (New York Meridian, 1974) Thomas Reid was one of the first to interpret Locke as defending a memory criterion, but his interpretation is not uncontested, see, for example, Marya Schechtman, *The Constitution of Selves* (Ithaca Cornell, 1996), pp 27-30

[3] "Of Personal Identity" [1736], in Perry, ed , pp 99-105

[4] Noonan, p 68

[5] See, for example, Parfit, p 220, Shoemaker, p 281

It is possible for us to have nonveridical memories as a consequence of, say, delusion, hypnosis, or having our brains tampered with We are capable of seemingly remembering events we never witnessed Obviously, such nonveridical memories have nothing to do with personal identity, I am not Napoleon when in one of my less clear moments I "remember" losing a battle at Waterloo In order for a specific memory to coconstitute personal identity, then, that memory is required to be veridical Establishing whether or not a memory is veridical means, however, establishing whether or not the person who remembers really witnessed the remembered event And that just means establishing whether the witness and the rememberer are identical Thus, personal identity seems to be presupposed by veridical memory so that, on pain of vicious circularity, the latter cannot be a criterion for the former

The circularity objection (in the general form in which I will discuss it here) says that the definition of my having a veridical memory is

(1) I am presently in a state as if remembering having a specific experience
(2) The content of that state is qualitatively identical or very similar to a past experience
(3) I had that experience

If this definition is correct, memory cannot be said—informatively—to coconstitute personal identity, since the identity between the person stage who had the original experience and the person stage who remembers having that experience is already presupposed by (3)

Shoemaker (*op cit*) was the first to come up with a substantial solution to the circularity problem In order to introduce his solution, let me discuss a peculiarity of the above definition

Suppose I had experience *E* (I saw a bear, say) but completely forgot about it Later, I am told by a friend about an event that is identical to the one that caused me to have *E* My friend did not witness any event of the sort, her story is made up By telling me her fiction, she unintentionally supplied me with every conceivable detail of an event I witnessed but forgot about The story triggers my imagination, endowing me with vivid images of the event I slowly begin to identify with these images and start to internalize them "Mistakenly," I begin to fit them into my biography The distinction between these images and my real memories begins to fade until I am finally under the impression that I really remember having had experience *E*

In the above case, (1), (2), and (3) are satisfied Yet we are not normally inclined to call this a case of experience memory The fact that I had the experience and the "recollection" is completely coincidental The mechanism by means of which I acquired knowledge of my past is absolutely unreliable, to put it mildly For by the same mechanism I could have had a "memory" of an event I never witnessed, in fact, that would have been much more likely

The reason why the above mechanism is unreliable is that my state of "recollection" is in no way dependent upon my actually having had the "recollected" experience Our concept of experience memory assigns an explanatory role to an original experience in the generating of a state of recollection thereof In other words, in order for me really to remember *E*, *E* should have been causally active in the production of my state of recollection Thus it seems we must add a fourth requirement to the definition of veridical memory

(4) My memory is causally dependent, in the right kind of way, on the past experience that is qualitatively identical or similar to my memory

Importantly, (4) does not mention 'I' Shoemaker's insight was that (4) safeguards what (3) was called in to secure That is, normally the fact that I had experience *E* secures my veridically remembering *E*, because in normal circumstances this guarantees a causal-explanatory connection between *E* and its recollection In other words, (3) normally secures (4), while of these two, (4) really matters The point of the above example is that in abnormal circumstances (3) does not secure (4) Shoemaker's (and Derek Parfit's (*op cit*, pp 219-22)) proposal, therefore, was simply to drop (3) in favor of (4) (1), (2), and (4) yield a satisfactory definition of veridical memory which does not mention 'I' and hence does not presuppose personal identity

One consequence of this solution is that, if there are ways of causally connecting an experience had by person *a* to a recollection of that experience had by person *b*, then *b* can be said to veridically "remember" *a*'s experience Since this is not in line with our ordinary use of the term 'memory', Shoemaker has introduced the notion of *Q(uasi)-memory* Person *b* Q-remembers *a*'s experience The crucial theoretical move here is that ordinary memory can be considered a subspecies of Q-memory, that is, the subclass of causally connected experiences and recollections in which both the experience and recollection are instantiated in the (numerically) same brain Since Q-memory does not require reference to persons in its definition ((1), (2), and (4)), neither does ordinary memory

II THE TROUBLE WITH Q-MEMORY

Q-memory is impossible to square with a holism of the mental, not just as a controversial functionalist doctrine, but also as an uncontested part of common-sense psychology—or so I argue A definition of memory in terms of (1), (2), and (4) needs to be supplemented with a principle that secures a level of coherence of a memory with its psychological context in virtue of which it maintains its contents and epistemic entitlements and hence its status as a memory

Parfit's example of Q-memory is that of Paul and Jane Paul has been to Venice and remembers seeing Palladio's San Giorgio Jane has never been to Venice But after a neurological operation in which some of Paul's brain traces are instilled in her brain, she finds herself endowed with what seems like a recollection of looking at San Giorgio from the point of view Paul actually saw it

Examples such as these are highly artificial in their simplicity[6] They concentrate on experiences and recollections that are describable more or less in abstraction from other mental contents By doing so they obscure serious difficulties with Q-remembering other persons' experiences that stem from the fact that most experiences and memories require an often rather elaborate context to make sense Let me give an example to illustrate this point

Suppose you have just taken your children to school You have clear memories of walking the familiar (to you) route in your neighborhood to their school You remember what your children looked like this day You remember thinking things such as "It's time to buy a new coat for Mary," and so on And you remember the familiar sight of the school building Now suppose these memories are instilled, by surgical means, say, in my brain I do not know your children and have never visited your neighborhood What would my Q-memories of your experiences be like?

I might have a picture in my head of your children walking the route they walked and entering their school But the route and the school will not evoke feelings of familiarity, and the children will seem strangers Marya Schechtman argues that, if I do not change my beliefs about the neighborhood and the children not being mine, my Q-recollection would be significantly different from yours, so that the similarity requirement between experience and recollection ((1) and (2)) is violated[7] If, on the other hand, I somehow get the belief instilled in me that the imagined neighborhood and the children are mine, the memory might resemble the original experience, but only at the cost of endowing me with a false belief about myself. I claim that the second of these possibilities is in fact no option, and that the first presents the Q-memory theorist with a problem more serious than the absence of qualitative similarity between experience and recollection might suggest.

---

[6] See also Schechtman, "Personhood and Personal Identity," this JOURNAL, LXXXVII, 2 (February 1990) 71-92, and "The Same and the Same Two Views of Psychological Continuity," in *American Philosophical Quarterly*, XXXI, 3 (1994) 199-212

[7] "Personhood and Personal Identity," pp 79-86

Let us start with the second possibility. Artificially endowing a person with the required background beliefs meets with a problem similar to the one it is supposed to solve. Just as we cannot endow someone with an isolated memory and expect its entire content to fit into any new psychological context, so we cannot endow a person with a belief just like that.

Suppose that along with the Q-memory I get the belief instilled in me that the children I Q-remember taking to school are *my* children. The fact that this single belief contradicts an immense number of memories and beliefs I have about myself would be more than sufficient for me to reject it as a delusion.[8] We might try to remedy this by endowing me with a large number of other background beliefs, and possibly even erase a few of my actual beliefs. But beliefs have a mind-to-world direction of fit. Hence my environment—my family, work, neighborhood, in short "my life"—will put rather severe constraints on what is rationally admissible for me to believe. It would be highly coincidental, then, if the background beliefs I am required to have did not clash with my environment (think only of the belief that the school I Q-remember is near to where I live). Similarly, it would be extremely unlikely that the actual beliefs of mine that should be erased because they contradict either the belief that the Q-remembered children are mine or the background beliefs that belief requires can indeed all be erased without severely impairing my rationality.

In many (quite possibly all) Q-memory cases, then, endowing a person with a Q-memory of someone else's experience will result in a clash between either this Q-memory or its required background beliefs with part of the new psychological context. The problem with Schechtman's second possibility is not only that the Q-"rememberer" will be endowed with false beliefs. The problem is that the incoherence of a Q-memory with its psychological context cannot be remedied by adding background beliefs; there will always be incoherence

---

[8] See, for example, Daniel Dennett, "Mechanism and Responsibility," in Gary Watson, ed., *Free Will* (New York: Oxford, 1982), pp. 168-69.

And that leaves us with the first possibility the Q-recollection (with or without added background beliefs) is more than likely to differ substantially from the original experience

If all a clash between a Q-memory and its new psychological context results in, however, is the fact that the Q-memory will not be qualitatively similar to the original experience with which it is causally connected, this need not yet deter the Q-memory theorist The qualitative similarity of the *complete* experience and recollection is not sacrosanct [9] It may easily be admitted that only part of a Q-recollection resembles the original experience due to the difference in psychological context of the two mental states In fact, if what seems like a memory does not entirely fit its psychological context, so to speak, this may be taken as an indication that one is Q-remembering someone else's experience, provided one knows such a thing to be possible [10]

But the displacement of a Q-memory in an "alien" psychological context also results in the loss of content and epistemic merits of a memory, or even the impossibility of there being something that resembles a memory state at all, depending on what theory of mind one adopts Let me elaborate on that

The worst-case scenario for the Q-memory theorist follows from the kind of holism with regard to mental content which is pervasive in most versions of what may well count as the best established theory of mind, functionalism, but also in such theories as interpretationism or anomalous monism [11] According to this holism, inferential and evidential interconnections between propositional attitudes (among which are the beliefs to which memories give rise) are coconstitutive of their contents If such holism were true, many cases of Q-memory would not just be cases in which the Q-memory differs qualitatively from the experience that caused it, rather, they would be cases in which there would be no similarity in content whatsoever between the newly inserted brain state and its causal origin. If inferential and evidential connections constitute content, then the clash between a Q-memory and its newly acquired psychological context will result in the Q-memory's having a nonsensical content or, rather, no content at all.

---

[9] It need not count against a theory of memory to drop the requirement that an experience memory should resemble an original experience in its entirety In fact, the requirement is too strong anyway If my recollection of seeing a bear last summer vacation was accompanied by the thought I had then—"I am seeing a bear now"—I am not remembering that event but having an illusion If, on the other hand, my memory is accompanied by the thought "I saw this bear so many months ago," I am having a memory of that event even though the contents of my mind do not completely resemble those at the time when I saw the bear See section IV

[10] See also John McDowell, "Reductionism and the First Person," in Jonathan Dancy, ed , *Reading Parfit* (Cambridge Blackwell, 1997), pp 240-41

[11] On "interpretationism," see, for example, Dennett, *The Intentional Stance* (Cambridge MIT, 1987), and on "anomalous monism," see, for example, Donald Davidson, "Mental Events," in *Essays on Actions and Events* (New York Oxford, 1980), pp 207-25

Moreover, to insert a Q-memory in a brain will be to tamper with the structure of its causal network. And since, according to the theories under consideration, propositional attitudes are global states of such a network, to tamper with this network is to distort not just the content of the inserted memory but also the contents of a whole host of other global states. If *constitutive holism* (as we may call it) is true, individual memories are not linked to their original experience by a separate causal connection. Rather, the whole frame of mind by which a recollection is coconstituted is linked to the whole frame of mind by which the original experience was coconstituted, without it being possible to single out the causal connections that link individual mental states. If constitutive holism is true, Q-memory is a myth.

But constitutive holism is not uncontroversial.[12] Theories of mind such as substance dualism, type-identity theory, or (versions of) computationalism,[13] are more inclined toward some kind of psychological atomism. Thus it seems they may allow for cases of Q-memory.

There is a much less strict mental holism, however, that is not part of a theory of mental content but rather part of our common-sense notion of how mental states interrelate, it is part of the explanandum rather than explanans, so to speak. This is the holism according to which a person can hold only so many incompatible beliefs before she can no longer be said to possess the rationality that is required for the ascription of beliefs and desires to be explanatorily profitable—even though such lack of rationality would not make mental-state ascriptions senseless, as they would be in the case of constitutive holism. It is the common-sense holist who says that I can believe that *X* only if I have the background knowledge required to make sense of '*X*' and to tell under what circumstances '*X*' is true or false. Such holism is in no way precluded by dualism, type-identity theory, or computationalism.

---

[12] See, for example, Jerry Fodor and Ernest Lepore, *Holism: A Shoppers Guide* (Cambridge: Blackwell, 1992).

[13] On "type identity," see, for example, J. J. C. Smart, "Sensations and Brain Processes," in V. C. Chappell, ed., *Philosophy of Mind* (Englewood Cliffs: Prentice-Hall, 1962), pp. 160-72, on "computationalism," see, for example, Fodor, *A Theory of Content and Other Essays* (Cambridge: MIT, 1990).

Common-sense holism does not oppose Q-memory the way constitutive holism does. But it does undermine the (epistemic) potential of Q-memory. Just as a Q-memory is likely to clash with part of its new psychological context, so it is likely that the new psychological context does not contain all the resources that are required for the Q-memory to retain its full content and epistemic entitlements. For instance, the richness of your experience in taking your children to school is lost on me. Moreover, due to the absence of relevant background knowledge, the information to which this Q-memory gives me access is very limited compared to your own recollection of the experience. For example, while you remember Jim and Mary entering their regular school (with its history in their lives, and so on), all I Q-remember is some kids entering a school-like building they might as well be seeing for the first time. Similarly, you may remember your daughter to be uncharacteristically sad that day. All I Q-remember is that there was a sad girl.

So, even if I recognize my Q-memory of your experience as a Q-memory and do not discard it as an illusion, it will not have all the epistemic entitlements it has to you. Importantly, this is due simply to the new psychological context, no causal connection can make up for the loss in epistemic value. To this, the Q-memory theorist may reply that, even if Q-memory is unlikely to carry as much information as ordinary memory, this does not mean that it carries no information at all. While Q-remembering your experience of walking your children to school, I may conclude that there were children who looked a certain way, who were walking toward a building that looked like a school while being accompanied by someone whose perspective I relive, and so on. Surely, my drawing these conclusions requires me to know that Q-memory is possible, but that is not specifically problematic.

Although correct, this reply misses a crucial point. The information carried by the Q-memory which is not lost on me is accessible in virtue of the *contingent* fact that my mental make-up is such that it allows me to access it. There is nothing in the causal connection between your experience and my Q-recollection that secures there being an appropriate context to access the information that is transmitted by it. If a Q-memory of your experience were instantiated in the brain of a member of some exotic tribe that has not had conscious contact with Western civilization, virtually all information contained in the memory would be lost on that person. The person would simply not know what she is Q-remembering, even if she knew that she is Q-remembering. (If this example is not convincing, we can always bring out the Martians, but I trust I shall not have to.)

It follows from common-sense holism, then, that the contents and epistemic merits of a (Q-) memory are just as much dependent upon the memory's psychological context as they are on the causes that produce the brain state upon which they supervene. Parfit's supposition—that only the causal origins of a memory count—presupposes a psychological atomism that can be sustained only by focusing exclusively on the phenomenal aspects of memories, as is done in the Paul and Jane case. But such atomism cannot be sustained when the conceptual side of memory is taken into account. Jane, for instance, can Q-remember Paul's experience only as an experience of Palladio's San Giorgio in Venice when she knows who Palladio was, what and where Venice is, and so on. And because the very contents of memories depend on their psychological context, so do their epistemic entitlements.

To conclude, on common-sense holism, (1), (2), and (4) do not make for a complete definition of memory. Apart from a causal connection between an experience and a recollection, an appropriate context for the recollection needs to be secured which allows the memory to retain some or most of its content and hence its epistemic merits. Part of such a context might be expected to exist in many cases of Q-memory, simply because, in general, people share a large portion of their background beliefs, if only because they inhabit the same world. But this is a contingent fact that is not in itself secured by the presence of the required causal connection between experience and Q-recollection.

Constitutive holism yields a similar, though stronger conclusion. On this theory, a causal connection between an individual experience and an individual recollection is also not sufficient for memory. Since neither the experience nor the recollection can be individuated in abstraction from its respective psychological context, a definition of experience memory requires a principle that secures the presence of the appropriate recollection context, given what the experience context was like.

What both kinds of holism imply is that a definition of memory in terms of (1), (2), and (4) needs to be supplemented by some kind of context-continuity requirement: a principle that explains and secures—rather than presupposes—the presence of a psychological context of a recollection which is related to the context of the original experience such that it allows the recollection to retain a sufficient portion of the original experience's content and hence have sufficiently enough epistemic merit to count as memory. Even without an answer to the difficult question of what counts as "sufficient" and "sufficiently enough," the incompleteness of Q-memory is plain.

In normal circumstances, the identity (or spatiotemporal continuity) of a body and brain secures (1), (2), (4) and the context-continuity requirement that cannot be captured by these three. This fuels the suspicion that memory defined in terms of (1), (2), and (4) in fact presupposes identity of the brain, or of the brain and body. Proponents of a bodily criterion of personal identity, but especially proponents of a brain criterion,[14] will find themselves in a rather good position to claim that, even on a causal definition of memory, memory does presuppose personal identity.

In what follows, I argue that a psychological criterion of personal identity need not be rejected on the basis of this circularity objection. What memory presupposes, and what fills in the details of the context-continuity requirement should be spelled out in terms of a type of psychological continuity rather than in terms of the unity of the brain or brain and body. This may seem to be an odd claim, given that psychological continuity itself has always been analyzed at least partly in terms of memory. I claim, however, that psychological continuity in the required sense is not constituted by memory, memory is an *epistemic*, rather than an *ontological* criterion for it.

III. TWO KINDS OF PSYCHOLOGICAL CONTINUITY

The circularity objection (in its contemporary guise) says that psychological continuity cannot analyze personal identity, since (A) psychological continuity consists of overlapping chains of psychological connectedness (in a sense to be explained below), and (B) the most important kind of psychological connectedness, memory, presupposes personal identity. The Q-memory solution to this objection attempts to deny (B). I argue against (A) that psychological continuity cannot be analyzed in terms of psychological connectedness.

Psychological continuity as I characterize it does not presuppose personal identity. Hence it can possibly analyze it—without circularity. Memory presupposes psychological continuity, I claim, for one thing because this provides the "context continuity" memory requires. Since psychological continuity might analyze personal identity, memory presupposes personal identity. But there is no harm done by that fact. For memory itself should be understood as a mere epistemic criterion of personal identity That, at least, is what I shall argue in an attempt to do justice to Butler's original point, to the contemporary circularity objection, and to the neo-Lockean insight that personal identity is constituted by psychological continuity

---

[14] On "bodily criterion" proponents, see, for example, Bernard Williams, "Personal Identity and Individuation," in *Proceedings of the Aristotelian Society*, LVII (1957): 229-52, and David Wiggins, *Sameness and Substance* (New York: Blackwell, 1980), on "brain criterion" proponents, see, for example, Thomas Nagel, *The View from Nowhere* (New York: Oxford, 1986).

But let me start by indicating how psychological connectedness (for example, memory) is supposed to constitute psychological continuity and what is wrong with this supposition

Detailed analyses of the concept of psychological continuity in the current debate are rare Parfit seems to articulate a near consensus on this issue when he defines psychological continuity in terms of overlapping chains of "strong connectedness" (*op cit*, pp 204-09) Some elucidation may be required here

Psychological connectedness denotes relations between mental states, such as those between experiences and recollections thereof, or between intentions and their being carried out, the continuous holding of beliefs and preferences, and the stable persistence of character traits It is easy to see why psychological continuity theorists emphasize such relations of qualitative similarity or identity The debate over personal identity is a debate over what makes two differently time-indexed person-stages stages of the same person When defining a psychological criterion of personal identity, what is more natural than to emphasize what remains the same, psychologically speaking, throughout large parts of a person's life?

Connectedness cannot analyze personal identity, however, for while identity is a transitive relation, connectedness obviously is not Take memory I may now remember my tenth birthday, and on my seventieth birthday I may remember writing this paper I may not, however, remember my tenth birthday on my seventieth Thus, by the criterion of memory, my tenth birthday-self and my seventieth birthday-self are unconnected, even though my present self is connected to both In order to remedy this deficit, Parfit introduced the notion of overlapping chains of strong connectedness

Two person-stages are stipulated to be strongly connected, if there are at least half the number of direct psychological connections between them which hold in the lives of every normal person over one day The concept of psychological continuity, now, is defined as the holding of overlapping chains of strong connectedness If chains of strongly connected psychological states continue to overlap in time (the way fibers in a rope overlap in space), there will be psychological continuity Thus, if enough direct psychological connections (memories, intentions, persistent beliefs, and desires, and the like) overlap in my life between my tenth and my seventieth birthday, these two person-stages will be continuous, even if no direct psychological connection holds between them

Characteristic of the Parfitian definition is that psychological continuity is exhaustively defined in terms of psychological connectedness The definition is hardly ever challenged explicitly But as an exhaustive definition, it just will not do It is absurd to claim that the multitude of mental states comprising a person's mental life is held together merely by relations of qualitative similarity and co-occurrence (in the same mind)

Consider Shoemaker "psychological continuity is constituted no more by [psychological connectedness] than it is by evolution and execution of plans of action, by deliberation and reasoning, and by countless other mental processes" (*op cit* (with Swinburne), pp 95-96) His point is to include psychological activities such as deliberating and thinking in the class of continuity-constituting connections I shall not be concerned merely with thinking and deliberating in what follows Rather, I shall elaborate on an implication of Shoemaker's point that continuity of our mental lives not only consists in relations of qualitative similarity between mental states (for example, memory), but also and importantly in relations between consecutive qualitatively nonsimilar states

Here, I want to concentrate on some of the conditions under which the succession of qualitatively distinct mental states yields a kind of psychological continuity that cannot be defined or analyzed in terms of Parfitian connectedness As a first abstract characterization of the continuity I am after, let me offer the following consecutive qualitatively distinct mental states can be continuous in virtue of the fact that their order of appearance crucially determines their full content and the intelligibility of their occurrence Differently put, a sequence of mental states can be continuous when the contents and intelligibility of later states in the sequence are codetermined by a diachronic context consisting of the earlier stages of the sequence Let me put some flesh on these abstract bones by means of some examples

Every stage in a thought process is usually intelligible only given its preceding stages To use a banal case after wondering what the product of 3 and 4 is, I realize it is 12 My thought "12" can only be understood as the answer to the preceding question in virtue of its place in the overall sequence of thoughts Likewise, the occurrence of "12" would be unintelligible for me if I were not aware of its place in the thought sequence in which it occurs It is the order in the sequence, not my being aware of it, that connects the question to the answer in a way that makes them continuous in the sense I am after

Thought processes and the like reflect what goes on "in" our minds There are, however, also important examples of sequences of mental states reflecting what goes on "outside" our minds and bodies (this is a divergence from Shoemaker, one that will turn out to be important) Take another example the contents of your mind when you are having a conversation with a friend If a sample of the contents of your mind after half an hour of conversation were transplanted to someone else's brain (for example, your hearing your friend say "I know, but does Julie feel the same about that"), that "receiver" would not be able to grasp the full content of that state (even if she knew its causal origins) and would be baffled about its occurrence You, on the other hand, are in a position to understand perfectly well what is being said and why, because your representation of the utterance occurs at a specific time in a sequence of representations "depicting" the whole of your conversation

Another example of a mental sequence reflecting external states of affairs is the consecutive occurrence of states representing sensory perceptions Sequences of perceptual contents cohere in virtue of the intelligibility of the world we perceive Each perception at a given moment is continuous with preceding and subsequent ones through the total intelligible "picture" of the surroundings of the perceiving body they draw Each movement of the body evokes a change of perception in accordance with the new position the body acquires Our thorough acquaintance with the basic laws of the three-dimensional world of medium-sized objects (including our body) allows us to retrieve from these consecutive impressions two interlocking stories that are already present in them one about the body's movements through space, and one about the perceptual features of the body's consecutive surroundings What constitutes the kind of continuity I am after is the fact that the sequence of contents coheres such that we can (though need not) retrieve these stories

It is a feature of this kind of continuity that what may seem to be two qualitatively identical mental contents on Parfit's atomistic, Humean bundle-theory of continuity, may well turn out to be different contents once the relevance of their diachronic mental contexts is recognized Take, for example, two qualitatively identical (for the sake of the argument) visual perceptions representing the Eiffel tower, had by two distinct persons One perception reflects a movie screen in London on which a picture of Paris is being projected, the other one reflects the Parisian construction itself The full contents of these representations differ, even if their visual/phenomenal aspects are identical The content of the impression of the person in London is that she is "looking at a picture projected in a theater in

London " This content is not derived from the impression alone It is derived from the impression's being embedded in a context—for example, walking through London, entering a cinema, watching a movie that contains a picture of the Eiffel tower Likewise, an impression caused directly by the real Eiffel tower is embedded in a context—for example, of visiting Paris as a tourist, wandering along the Seine, and so on Each of the two contents is codefined, so to speak, by their diachronic context

In order to have a label by means of which I can refer to these progressions of psychological contents, I shall use the term *narrative continuity*[15] (N-continuity for short) with a number of qualifications 'Narrativity' is not supposed to introduce teleology or constructivism into the notion of psychological continuity Nor is the term intended to suggest that psychological continuity involves a plot or a "point", the stories involved are usually of a mere "and then , and then " kind The aspect of narrativity that nevertheless renders the term apt is the one Alasdair MacIntyre[16] drew attention to "a perception" or "a thought"—for instance—taken in isolation is an abstraction from a whole perceptual sequence or a complete process of thought Just as a particular event in a story acquires its full sense only as part of the whole story, individual thoughts acquire their full meaning only as part of a process of deliberation, and individual sense perceptions acquire their full sense only as part of a sequence of perceptions portraying a body's movements through space-time It is the whole of such a sequence that endows its "parts"—which are abstractions rather than constituents—with their full meaning

The notion of N-continuity as described so far is still relatively vague It is extremely tricky to try and offer a formal definition of it, however [17] Instead, let me defend four claims about N-continuity which will highlight those aspects of it which will be relevant to the overall argument

(1) N-continuity does not require or presuppose memory
(2) N-continuity is well capable of capturing the transitivity that is a salient feature of identity by providing a life-long psychological

---

[15] No similarity with Paul Ricoeur's narrative theory of personal identity is intended, see *Oneself as Another* (Chicago University Press, 1992)

[16] *After Virtue* (Notre Dame University Press, 1981) pp 217ff

[17] See Igor Douven, "Marc Slors on Personal Identity," *Philosophical Explorations*, II, 2 (1999) 143-49, and my "A Reply to Igor Douven," *Philosophical Explorations*, II, 2 (1999) 150-52 For a more elaborate treatment of N-continuity, see my *The Diachronic Mind An Essay on Personal Identity, Psychological Continuity and the Mind-Body Problem* (Boston Kluwer, 2001), chapter 3

narrative In normal circumstances, in the actual world, this is due to the fact that we are embodied beings

(3) Nevertheless, N-continuity is not bodily continuity in disguise

(4) Although N-continuity is more than likely to be causally underpinned, it cannot be analyzed in terms of causality

If these claims can be established, N-continuity can be considered a kind of psychological continuity which is distinct from Parfitian continuity, notably due to the fact that it does not involve memory, but which is potentially just as much capable of analyzing personal identity

(1) *N-continuity does not presuppose memory* I have been talking so far about the relevance of the diachronic context of a mental state in determining this state's content and intelligibility, without distinguishing between the past mental states that make up this context and their representations that (usually) co-occur with the mental state whose content they codetermine The distinction is important, however, for it may seem as if N-continuity involves memory, and the distinction helps to explain why this is not the case To use some labels, let me speak of the *diachronic* context of a mental state when I mean to refer to earlier mental states in a sequence, and speak of *cotemporal* context when I mean to refer to the set of co-occurring mental states that may or may not contain representations of (part of) the diachronic context

The idea that N-continuity presupposes memory is the idea that for the full content and intelligibility of a given mental state *m* to be codetermined by that state's diachronic context, this context (or at least the relevant bits thereof) ought to be represented in the cotemporal context of *m* Since the representation of earlier mental states in a present frame of mind just is memory, if representation of the diachronic context in the cotemporal context is required for there to be N-continuity of *m* with its mental predecessors, this would mean that N-continuity requires memory

In order to keep N-continuity and memory apart, it is crucial to distinguish sharply between N-continuity, on the one hand, and awareness thereof, on the other While awareness of N-continuity of *m* with its predecessors does require the diachronic context of *m* to be represented in its cotemporal context, N-continuity itself does not require this That is, the representation of earlier mental states is required for a "subject"[18] to be aware of, or have access to, the full content of her mental states But the full content of a mental state in the sense relevant to N-continuity is not determined by her actually accessing it, it is codetermined by her being able to do so, that is, by the occurrence of previous mental states, not by these previous states' later representation

---

[18] The term 'subject' is not meant to refer to a substantial self that exists independently of the mental states it 'has' Rather, it refers to a function of these

An example might help here Take the above case of the two perceptual representations of the Eiffel tower, one caused by the tower itself, and one caused by a picture on a movie screen Now suppose that the two persons in these perceptual states suffer from amnesia Their long-term and short-term (episodic and iconic) memory are impaired (their semantic memory can be left intact) One perception is that of the real Eiffel tower in Paris, the other is of a picture on a movie screen in London Given their amnesia, however, neither of the two persons is capable of accessing these full contents From their perspectives, both states have identical contents But that is merely because they both have access only to the visual/phenomenal part of their perceptual states Given their amnesia, they do not have access to the (N-continuous) stream of mental states which leads up to these perceptions and endows them with contents richer than their phenomenology can be credited for But that does not make these perceptions less N-continuous with their preceding mental states they still have a content that is there to be potentially discerned and put to use as soon as the amnesia is over, a content that is determined by their diachronic psychological context

One might object to the term 'content' when a person need not be aware of it But such an objection would not refute the dissociation of N-continuity and memory Accommodating it can be done in either of two ways One is to redefine N-continuity in counterfactual terms one mental state *m* is N-continuous with a series of previous mental states if and only if when *m*'s cotemporal psychological context contains representations of these previous states, *m*'s full content becomes available to the subject Alternatively, the notion of 'content' itself can be cast in counterfactual terms A state's full content, then, is determined by what would be available to the subject should the diachronic context of this state be represented in its cotemporal context The state would have this content, even when the diachronic context is not actually cotemporally represented This position would be analogous with externalism about mental content Here, too, a

---

states themselves, such as the having of a consistent set of second-order states The subject, then, is something like a center of narrative gravity, to use Dennett's term state can have its content determined by things that are not necessarily available to the subject

Either way, a given mental state can be N-continuous regardless of whether its cotemporal context actually contains representations of previous mental states A mental state can thus be N-continuous with its predecessors even though a lack of memory precludes awareness of this The fact that N-continuity and memory are correlated more often than not is just as well, though, since it allows us to access the full contents of our mental states But the correlation is contingent—N-continuity itself does not require memory

(2) *The N-continuity-based unity of our psychological lives* In labeling the mental processes I am concerned with here cases of narrative *continuity,* I am making an as yet unargued for, but nevertheless important assumption The term 'continuity' is meant to refer to a relation that is characterized by the transitivity that is a salient feature of identity but not of connectedness For narrative sequences to be called "continuity," it must be shown that such sequences constitute a life-long transitive relation other than merely by overlapping In other words, there has to be a sense in which the mental life of one person consists of one narrative sequence

There is a sense in which my mental life is one large narrative, despite the fact that my thoughts may jump from topic to unrelated topic, despite the fact that feelings can sometimes arise "out of the blue," despite the fact that I lose consciousness when I sleep, despite—to summarize—the fact that our consciousness is patchy like a quilt and (in Daniel Dennett's term) "gappy " Let me first elaborate on how gappiness and patchiness does not stand in the way of there being one narrative organization of the mental lives of actual persons, and then turn to the question whether such an organization can survive hypothetical scenarios such as brain transplants

With regard to our actual situation, what I have in mind is this despite the fact that our psychological lives are gappy and patchy, there is a *basic* narrative that is neither patchy nor gappy This basic narrative is represented by our consecutive perceptual contents No doubt other types of content also play a unifying role, but I concentrate on perceptual contents because their unifying role is, I think, more transparent than that of other types

The two stories told by the successive impressions of our senses are, as I briefly indicated, the complementary stories of a body's movements through space and time and the story of the particular perceptible features of its surroundings at consecutive times Both stories interlock into one story about an objective continuant's successive whereabouts Successive perceptions are the subjective counterpart of the objective event of one body moving through space and perceiving its surroundings.

Successive perceptions acquire narrative coherence in virtue of the fact that they are caused by one body's movements through a stable (not static) physical world. We are acquainted with the character and proceedings of this world so that we are able to appreciate this coherence. This is largely in agreement with John McDowell: "continuous 'consciousness' is intelligible (even 'from within') only as a subjective angle on something that has more to it than the subjective angle reveals: namely, the career of an objective continuant." (*op. cit.*, p. 233).

As it stands, though, McDowell's claim is stronger than what I am after. For one thing, I do not want to claim that psychological continuity as a whole hinges on the continuity of an objective continuant or body. I only claim that N-continuity of perceptual contents does. The narrative unity that perceptual contents provide, I shall claim, is necessary but not sufficient for full-blown psychological continuity. Furthermore, I do not take it as a conceptual truth that the diachronic unity of mind that constitutes personal identity over time—according to a psychological criterion—requires a perceptual narrative for which bodily continuity, in turn, is a prerequisite. It may well be an empirical fact, though, that the diachronic unity of mind of all actual persons is secured for a large part through bodily continuity.

With these qualifications in mind, and bearing in mind that we are discussing actually existing persons, the crucial unifying role of perceptual contents stems from the fact that, in Peter Strawson's[19] words, "a temporally extended series of experiences should have a certain character of connectedness and unity, secured to it by concepts of the objective. That experience should be experience of a unified objective world at least makes room for the idea of *one* subjective or experiential route through the world, traced out by *one* series of experiences which together yield *one* unified experience of the world—a potential autobiography" (*ibid.*, p. 193).

Perceptual contents accompany virtually all of our conscious life. Given that its source, an intelligible world, is a coherent whole—space-time and causally related objects—their sequences cannot be patchy. Finding myself at place *a*, I need not expect to have perceptions of place *b* 100 miles away within the next ten minutes. We cannot be here one moment and a hundred miles from here the next. If we go from *a* to *b*, we shall pass through all intermediate places separating *a* and *b* according to the route we take, thus producing a sequence of perceptions narrating the story of this route (the position of our bodies at different places, the position of our eyes, and so on).[20] The story simply cannot just stop to make room for a different one. Perceptual continuity, then, cannot be patchy. (And if it appears so, we know we are the victim of an illusion, the uncovering of which restores the continuous story.)

---

[19] *The Bounds of Sense* (London: Methuen, 1966), see also Quassim Cassam, "Kant and Reductionism," *The Review of Metaphysics*, XLIII (1989): 72-106, "Parfit on Persons," in *Proceedings of the Aristotelian Society*, XCIII (1993): 17-37, and Gareth Evans, *The Varieties of Reference*, McDowell, ed. (New York: Oxford, 1982).

Perceptual continuity is gappy, though. Every now and then, mostly when we go to sleep, we lose consciousness. But this does not at all imply that the narratives that are conveyed by or implicitly present in our perceptions are gappy. Usually, we wake up where we went to sleep. The story of our body's movements, then, continues where it was broken off. It is not unimaginable, though, that we wake up somewhere else. In such a case, it is misleading to say that the perceptual narrative is gappy or incomprehensible. The concept of our body as a physical object—which may either be the product of induction from earlier perceptions, or of evolutionary hardwiring—regulates the story told by our consecutive perceptions, whether we are aware of it or not. Thus, the diachronic context of one's perceptions when waking up somewhere else than where one went to sleep, will endow those perceptions with a content that, in linguistic terms, reads "the perceptions while waking up after the body that issues them has been moved while unconscious," or something to that effect. Perceptual narratives, then, are subjective counterparts of complete objective narratives, and as such not gappy.[21]

N-continuous perceptual contents provide our psychological lives with what I shall call a *basic narrative*. A perceptual narrative is "basic" in the sense that it provides a lifelong narrative that unifies the multitude of mental states that makes up a "stream of consciousness," while it is in itself insufficient for full-blown psychological continuity. It is an elementary thread running through my "stream of consciousness," connecting all its different temporal stages. Moreover, it provides an elementary "format" for the structure of such streams by forming the diachronic background against which all and only the mental states of one person occur Such a background puts rather severe constraints on thoughts, intentions, beliefs, and so on, thus securing a high degree of coherence, since most nonperceptual mental states are specifically connected with and shaped by a particular episode of a basic narrative Let me elaborate on that

---

[20] Parfit's teletransportation thought experiment seems to contradict this. For reasons to ignore teletransportation, see the last section of my "Two Conceptions of Psychological Continuity," *Philosophical Explorations*, I, 1 (1998): 59-78.

[21] This does not preclude our misconstruing perceptual narratives or our misrepresenting them when we remember (part of) our past narratives. The point here is merely that *any* interpretation of a string of perceptions requires the supposition that there is at least an objective story to be told which connects these perceptions with a certain well-determined set of sense organs.

Nonperceptual mental contents such as beliefs and thoughts do not merely co-occur with perceptual ones, as, for example, Parfit (*op cit*, pp 250-51) would have it Thoughts, desires, and beliefs may, for instance, be directed at items in a person's direct proximity, I may attend to what my conversation partner is saying, my attention may be with a book I am reading, I may be busy trying to find the quickest way of getting somewhere, I may be enjoying a concert, and so on, and so on In such cases, it is easy to see how thoughts interrelate with the basic narrative against the background of which they occur The same goes for intentions, which take into account our present whereabouts and relate that to a future action at a future place

But even when thoughts, beliefs, and desires do not concern a person's actions or direct environment, there are almost always conceptual, evidential, or biographical ties to the subject's basic narrative Conceptual ties stem, for example, from our acquaintance with a stable physical world—including our bodies—and its basic regularities, which is conveyed through our basic narrative (My belief that the Eiffel tower is in Paris implies that it is about 600 kilometers from where I am now, that it will take me about 5 hours by train to get there, and so on) Evidential connections between beliefs and a basic narrative are there simply because whenever there is perceptual evidence for a belief, a perception will also fit in and be part of a larger basic narrative Biographical connections between one's beliefs and desires and one's basic narrative, finally, stem from the fact that a basic narrative, being a major constituent of a biography, encompasses one's experiences and hence determines one's "food for thought "

Thus, a basic narrative is intertwined with most of our other thoughts, beliefs, and desires This secures a high degree of synchronic and diachronic coherence between these thoughts, beliefs, and desires themselves Cotemporal mental states of one person are subject to the same constraints issued by the episode of the basic narrative with which they co-occur This gives them a nonaccidental "place" on the "time-scale" of the basic narrative, so to speak, so that, otherwise unrelated thoughts, for example, nevertheless do have a rationale for their order of occurrence that confers to them a minimal degree of N-continuity. Since a basic narrative is life-long (starting when one's perceptual-cognitive apparatus is well developed), it can hence unify a psychological life into one N-continuous sequence.

(3) *N-continuity is not bodily continuity in disguise.* In the case of actually existing persons, my claim is that bodily continuity provides a basic narrative that unifies our psychological lives into one life-long narrative. Nevertheless, N-continuity does not simply boil down to bodily continuity supplemented with a number of psychological subnarratives.

While a basic narrative is necessary to secure a degree of narrative unity of a person's psychological life, and while bodily continuity is sufficient to provide a basic narrative given a properly functioning perceptual and cognitive system, nothing that has been said so far implies that bodily continuity is logically necessary for a basic narrative. That is, bodily continuity is very likely to be the only provider of a basic narrative, although we cannot rule out on a priori grounds hypothetical scenarios in which persons have full-blown N-continuity without bodily continuity.

Take a case in which the contents of a person's brain are transplanted into another brain and body. If bodily continuity is a necessary prerequisite for a basic narrative, then in such a case there would be a breach in narrative psychological continuity. We can imagine the case, however, such that the role of the basic narrative is provided—for a brief period anyway—by some other psychological mechanism than perceptual representations of bodily whereabouts. For instance, it is far from unimaginable, that, if the person were informed beforehand of the body-swapping operation, her expecting a radical change in proprioceptual and perceptual contents after the operation would "bridge" at least part of the narrative gap that threatens to disrupt N-continuity. Likewise, informing the person's peers about the operation might result in their confirming the body swap and hence the N-continuity of the person at issue, a confirmation that might make the difference between confusion or even mental disintegration, on the one hand, and mental coherence and sanity, on the other.

We can extrapolate from this example to the truly bizarre scenarios that turn up every now and then in the debate on personal identity. Take a scenario in which persons change bodies every week (or day, or hour, or minute). In such a scenario, we can imagine these persons being informed about these changes so that they are able to bridge what to us would be significant gaps in N-continuity. But we can also, and perhaps better imagine this information being "wired into" their cognitive systems, in much the same way that bodily continuity is wired into ours as the rationale for processing successive perceptual contents. In either case, these persons do have a basic narrative that unifies their psychological lives. But this basic narrative will provide a "format" for the structuring of the psychological lives of these persons which will differ substantially from ours. (For instance, while we can base our beliefs about the color of our hair right now on memories of looking into the mirror weeks ago, they cannot.)

To summarize, a basic narrative is a necessary condition for full-blown N-continuity. But since bodily continuity is only sufficient (given a properly functioning cognitive system) for such a basic narrative and not logically necessary, N-continuity is strictly a form of psychological continuity. It is not bodily continuity in disguise, despite the fact that in our actual world N-continuity does require bodily continuity.

(4) *N-continuity cannot be analyzed in terms of causality.* So far, the relations between mental states comprising an N-continuous stream are characterized purely in terms of the contents of these states, and not in terms of causal relations between them. This does not mean that N-continuity is likely to exist without being "causally underpinned." No plausible metaphysics of the mental would allow for the absence of causal connections—at least at the level of the (physical) realizers of mental states.

So far so good. But now the following question arises: in Parfitian psychological continuity, causality coanalyzes what continuity is about (continuity is overlapping connectedness, and relations of connectedness, such as memory, are codefined in terms of causality). Is it the same with N-continuity? The answer to this question is negative. The reason is that we cannot identify the causal relations underlying N-continuity in abstraction from N-continuity. The causal relations underlying N-continuity are way too complex and scattered for this, infinitely more so than the allegedly straightforward causal relations that secure psychological connectedness on the Parfitian model. There are two reasons for this.

First, there is a diachronic mental holism to which N-continuity is committed. The fundamental idea behind N-continuity is that mental states at one point in time acquire their full content in virtue of being embedded in a specific diachronic psychological context. Furthermore, we have to do justice to a version of the synchronic mental holism that inspired the argument against Q-memory (section II, it is unimportant which version). The "picture" of the mind yielded by both kinds of holism is one in which it is impossible to single out a given causal connection as the one that connects, for example, this experience with this recollection, or this question with this answer, or and so on. Such mental connections require what I earlier labeled *context continuity* And that means that a very large number of causal connections are at play in what in mental terms would be a single connection

Second, many causal connections underlying N-continuity do not exist wholly "within the subject" The best example here is the perceptual continuity involved in our basic narratives When I turn my head, I have consecutive perceptions of my computer, some books, a photo on the wall, a table, and so on These perceptions are N-continuous But they are not causally continuous in any sense that resembles the causal continuity involved in Parfitian psychological continuity One perception does not cause the next one

Nevertheless, causal relations do underlie the N-continuity of perceptual contents At consecutive times, light rays reflected from the objects of my perceptions cause changes on the surface of my retina which cause changes in the optic nerve, which cause certain perceptual states of the brain, the movement of my head causes my eyes to be in the positions in which they are susceptible to the light rays emitted from these objects, and so on But the point here is that these causal relations are scattered, so to speak They exist partly within the subject and partly without

The complexity of the causal relations underlying N-continuity, and the fact that many of them involve the outside world just as well as what goes on inside our heads, makes for a situation in which we can only identify the causal relations involved (in principle, not in practice) when we already have a firm grasp of the concept of N-continuity Without the notion of N-continuity, there is no simple principle by means of which to demarcate the class of N-continuity underlying causal relations (such as, for example, 'only the causal relations inside a person's head count') Hence, the causal connections at play cannot be identified in abstraction from N-continuity, so that the former cannot analyze the latter

IV MEMORY, N-CONTINUITY, AND THE CONTEXT-CONTINUITY REQUIREMENT

In the previous section, I argued that N-continuity does not presuppose memory, even though awareness of it does Now I argue that the relation between memory and N-continuity is asymmetrical memory does require N-continuity The reason is that N-continuity explains how and why the context-continuity requirement for veridical memories can be met, whereas it could not be met by a Q-memory based conception of psychological continuity (section II) There are at least three interrelated ways in which N-continuity is involved in this requirement such that it is presupposed by memory

(1) Each (experience-)memory requires an appropriate context to bring out its full content and epistemic entitlements, that is what the holism-based criticism of Q-memory shows Such a context consists, to a large extent, of other memories, ranging from concrete experience memories conveying the context in which one had the experience whose recollection is at issue, or more conflated, condensed, or summarized memories,[22] or even inferred (but not really remembered) knowledge of one's past For instance, the cotemporal context of my remembering my wife's last birthday dinner, contains concrete episodic memories (for example, of my cooking it), semantic memories conveying the meaning of, for example, 'birthday', 'wife', and so on, and condensed biographical memories conveying, for example, the fact that this birthday dinner is part of a tradition of birthday dinners The mechanism by which such contexts are established, then, is memory (and sometimes inference) and not N-continuity But such mechanisms are insufficient to explain why and how the context-continuity requirement is fulfilled

What needs to be explained as well is how these mechanisms get the appropriate "inputs," so to speak The context of a given memory consists, to a large extent, of the representation of bits of a person's biography that codetermine the specific content of the memory For a memory to have the right kind of cotemporal context, then, the remembered experience and the memory should be part of the same "psychobiography" that contains the experiences and knowledge to be potentially represented in the context of that memory In a largely atomistic conception of psychological continuity, there is no guarantee at all that a person's psychobiography contains the right kind of knowledge and experiences to make up a given recollection's context that allows us to make sense of the experience of remembering If, by contrast, the succession of mental states in a person's life is considered to be structured along the lines of N-continuity, yielding the psychological counterpart of a coherent biography, it is more than likely that for almost each memory a person has, there is a set of past experiences, bits of knowledge, and so on in the rememberer's psychobiography that, when represented as the memory's cotemporal context, allows that memory's full content to be accessible so that it retains its epistemic entitlements The fact that the actual representing is done via memory does not make N-continuity dispensable or redundant in an explanation of how the context-continuity requirement can be fulfilled

---

[22] See, for example, Schechtman, "The Truth about Memory," *Philosophical Psychology*, vii, 1 (1994) 3-18, C R Barclay and P A DeCooke, "Ordinary Everyday Memories Some of the Things of Which Selves Are Made," in U Neisser and E Winograd, eds , *Remembering Reconsidered Ecological and Traditional Approaches to the Study of Memory* (New York Cambridge, 1988), pp 91-126, L W Barsalou, "The Content and Organization of Autobiographical Memories," in Neisser and Winograd, eds , pp 193-244

(2) The full content of an experience-memory includes the "place" of the remembered experience on a "subjective time-scale " With the passing of time, the propositional contents of the beliefs involved in experience and memory change like this "It is *Y* now" (original experience), "It was *Y* a moment ago" (recollection shortly after the experience), "It was *Y* a while back" (recollection long after the experience), and so on [23] These contents have to change in order to keep one identical memory alive And they must change in an order that fits the (basic) narrative of the person's temporally distancing herself from the original event of experiencing (A similar point can be made about intentions [24])

For such a diachronic change of propositional contents to be possible, the context of a memory should contain some notion of the temporal distance between the present person-stage and the person-stage that had the original experience And for that it is crucial that both the experience and the recollection are part of a psychological narrative that is unified by a basic narrative For only thus can we conceive a psychological mechanism by means of which their temporal distance is represented with a (usually not too high) degree of accuracy the remembered experience is "located" on the subjective biographical time-scale of the rememberer's N-continuous psychological life (by relating it, for instance, to a number of time-marking life events) The propositional contents of memories thus require the corepresentation, albeit extremely abbreviated, of a part of the rememberer's N-continuous psychobiography in which past mental events have a definite order (This is in itself a form of N-continuity a memory requires a diachronic context for its full content)

(3) The context-continuity requirement hinges on the coherence of a memory with its psychological context This context contains both other memories and, for example, representations of the rememberer's physical and social surroundings But coherence itself requires norms The clash of a Q-memory with an alien context is the violation of some of these norms Conversely, the coherence of a given memory with knowledge of one's biography, as well as with one's present representations of one's objective and social environ-

---

[23] See Evans, *The Varieties of Reference*, pp 236-39

[24] See, for example, Shoemaker (with Swinburne), p 95

ment, is compliance with such norms So where do these norms come from?

A significant part of the answer to this question is that the elements that structure our basic narratives—continuity of the body as an object in the physical and social world, acquaintance with the behavior of medium-sized objects and their restrictions, and so on—are exactly those which determine a large part of the norms governing the coherence of a memory with its psychological context Why does your memory of taking your children to school (see section II) not fit into the context of my mind? Because, for example, my "internalization" of that memory would violate in various ways the norms that govern our basic narratives and the N-continuous psychobiographies that occur against their backgrounds There is no way in which my body could have traced a spatiotemporal path that would have endowed me, for instance, with your memory but not with memories telling me how and why I have wound up at this strange place taking two strange children to school, but with other memories (the ones I actually have) instead N-continuity and especially the basic narrative provides a relatively specific "format" for the organization of information about one's life which puts heavy restrictions on which memories can possibly represent experiences that would fit into a past stage of one's psychobiography

Since coherence is without a doubt the most crucial epistemic criterion for the veridicality of our memories, N-continuity including a basic narrative is a *conditio sine qua non* for memory as an epistemic device that gives us access to past experiences

V A NEW SOLUTION TO THE CIRCULARITY PROBLEM

The circularity objection against the idea that psychological continuity can analyze the concept of personal identity comprises two claims

(A) Psychological continuity can be analyzed in terms of overlapping chains of connectedness

(B) The most important kind of connectedness—veridical memory—presupposes personal identity because it can be defined thus

  (1) I am presently in a state as if remembering having a specific experience

  (2) The content of that state is qualitatively identical or very similar to a past experience

  (3) I had that experience

Requirement (B)(3) introduces personal identity into the definition of veridical memory Therefore, veridical memory presupposes what it means to coconstitute according to (A) Hence there is circularity in analyzing personal identity in terms of psychological continuity

The Q-memory solution attacks claim (B) by inventing a definition of memory that does not presuppose personal identity It does so by replacing (3) with

> (4) My memory is causally dependent, in the right kind of way, on the past experience that is qualitatively identical or similar to my memory

In section II, I argued that this "solution" is not, in fact, feasible A definition of memory in terms of (1), (2), and (4) crucially lacks a principle that secures the presence of an appropriate psychological context for every recollection, a context that potentially brings out its full content and epistemic entitlements

In section III, I sketched the outlines of a type of psychological continuity that cannot be analyzed in terms of overlapping chains of psychological connectedness N-continuity In section IV, I argued that N-continuity is a necessary condition for recollections to have appropriate psychological contexts to bring out their full contents and epistemic entitlements

This may seem to open up the possibility of patching up the Q-memory solution, (1), (2), (4), and N-continuity make for a possible definition of memory that does not presuppose personal identity In fact, the N-continuity of an experience and a recollection obviates the need for (4) in about the same way (4) obviates the need for (3) (see section I) Under any plausible theory of mind, the N-continuity of an experience and a recollection encompasses their causal connection We may therefore replace (4) with

> (5) My state of apparent remembering must be N-continuous with that past experience

(1), (2), and (5) can arguably make for a definition of memory that does not presuppose personal identity and does not fall prey to the objection leveled against the Q-memory definition

Once N-continuity is acknowledged, however, a much simpler solution to the circularity objection becomes feasible, one that explains (rather than explain away) the nagging intuition that memory cannot really be defined without personal identity Once N-continuity is acknowledged, (A) can be rejected psychological continuity cannot be analyzed in terms of overlapping chains of (strong) connectedness If (A) can be denied, the possible truth of (B) seems irrelevant My claim is that N-continuity may analyze personal identity noncircularly, N-continuity is a transitive, life-long kind of psychological continuity that does not presuppose personal identity the way continuity defined in terms of connectedness appears to

Where does this leave memory? Does memory become irrelevant to personal identity? My answer to this last question is both "yes" and "no "

I argued in section III that N-continuity does not require memory, but that awareness of it does If memory is required for us to be aware of N-continuity, and if N-continuity can analyze personal identity, then memory is required for us to be aware of our own identity over time Memory can thus be an epistemological criterion for personal identity without constituting it This is perfectly in line with Butler's original objection against Locke's memory criterion, which was meant to uncover Locke's conflation of epistemological and metaphysical criteria Recognizing this, however, does not imply the rejection of a metaphysical-cum-semantic psychological criterion of personal identity[25]

With regard to the Lockean psychological criterion, it is sometimes remarked that it contains a "benevolent" kind of circularity that cannot be argued away[26] The definition of memory in terms of (1), (2), and (5) still contains that circularity For the very concept of memory, which is an epistemological criterion for N-continuity/personal identity presupposes the concept of N-continuity/personal identity But this is no longer the vicious circularity that undermined the memory criterion of personal identity For N-continuity is not codefined in terms of memory, so that memory does not presuppose itself

MARC SLORS

Nijmegen University

---

[25] The proposed criterion of personal identity is only partly Lockean in spirit It is Lockean because it is a psychological criterion But Locke's criterion was thoroughly driven by empiricist concerns, personal identity could not, according to Locke, be something that we might not even be conscious of N-continuity, by contrast, is something we are often not aware of at all

[26] For instance, by McDowell in "Reductionism and the First Person", and by Shoemaker in his "Personal Identity and Memory," this JOURNAL, LVI, 22 (October 22, 1959) 868-82

NEW BOOKS ANTHOLOGIES

ANTON, ANATOLE, MILTON FISK, and NANCY HOLMSTROM, eds *Not for Sale In Defense of Public Goods* Boulder, CO Westview, 2000 xx, 468 p Paper $25 00

BAYLE, PIERRE *Political Writings* SALLY L JENKINSON, ed New York Cambridge, 2000 lxiii, 367 p Cloth $69 95, paper $24 95

BRAND, PEG ZEGLIN, ed *Beauty Matters* Bloomington Indiana UP, 2000 xv, 329 p Cloth $45 00, paper $19 95

BRANDOM, ROBERT B , ed *Rorty and His Critics* Malden, MA Blackwell, 2000 xx, 410 p Cloth $64 95, paper $26 95

BROGAN, WALTER, and JAMES RISSER, eds *American Continental Philosophy A Reader* Bloomington Indiana UP, 2000 viii, 396 p Cloth $45 00, paper $19 95

BRYDEN, MARY *Deleuze and Religion* New York Routledge, 2001 xi, 208 p Paper $24 95

BUCHWALD, JED Z , and I BERNARD COHEN, eds *Isaac Newton's Natural Philosophy* Cambridge MIT, 2000 xx, 354 p Cloth $45 00

BURNS, TONY, and IAN FRASER *The Hegel-Marx Connection* New York St Martin's, 2000 viii, 258 p Cloth $69 95

BURTON, STEVEN J The Path of the Law *and Its Influence The Legacy of Oliver Wendell Holmes, Jr* New York Cambridge, 2000 xi, 354 p Cloth $64 95

CHAPUIS, ANDRÉ, and ANIL GUPTA, eds *Circularity, Definition, and Truth* Atascardero, CA Indian Council of Philosophical Research, 2000 v, 402 p Cloth

COLEMAN, JULES, ed *Readings in the Philosophy of Law* New York Garland, 1999 680 p Paper $29 95

CRARY, ALICE, and RUPERT READ, eds *The New Wittgenstein* New York Routledge, 2000 ix, 403 p Paper $29 99

CRISP, ROGER, and BRAD HOOKER, eds *Well-Being and Morality Essays in Honour of James Griffin* New York Oxford, 2000 xi, 316 p Cloth $60 00

DAVIS, STEVEN, ed *Color Perception Philosophical, Psychological, Artistic, and Computational Perspectives* New York Oxford, 2000 vi, 247 p Cloth $65 00, paper $35 00

FAULCONER, JAMES E , and MARK A WRATHALL, eds *Appropriating Heidegger* New York Cambridge, 2000 x, 214 p Cloth $59 95

FOLLESDAL, DAGFINN, ed *Naturalism and Ethics Philosophy of Quine, Volume 2* New York Garland, 2001 xii, 400 p Cloth $135 00

FULLER, GARY, ROBERT STECKER, and JOHN P WRIGHT, eds *John Locke An Essay Concerning Human Understanding* New York Routledge, 2000 vi, 282 p Paper $85 00

GAUT, BERYS, and DOMINIC MCIVER LOPES, eds *The Routledge Companion to Aesthetics* New York Routledge, 2001 xviii, 581 p Cloth $100 00

GRACIA, JORGE J E , and PABLO DE GREIFF, eds *Hispanics/Latinos in the United States Ethnicity, Race, and Rights* New York Routledge, 2000 281 p Cloth $80 00, paper $22 99

HAHN, LEWIS EDWIN, ed *Perspectives on Habermas* Chicago Open Court, 2000 xiv, 586 p Paper $29 95

HARTSHORNE, CHARLES, and WILLIAM L REESE, eds *Philosophers Speak of God* Amherst, NY Humanity, 2000 xvi, 535 p Paper $27 00

HEYES, CECILIA, and LUDWIG HUBER, eds *The Evolution of Cognition* Cambridge MIT, 2000 ix, 386 p Cloth $47 95

HOAGLAND, SARAH LUCIA, and FAYE, MARILYN, eds *Feminist Interpretation of Mary Daly* University Park Penn State UP, 2000 xi, 452 p Cloth $70 00, paper $25 00

JAMIESON, DALE, ed *A Companion to Environmental Philosophy* Malden, MA Blackwell, 2001 xvi, 531 p Cloth $124 95

KARP, IVAN, and D A MASOLO *African Philosophy as Cultural Inquiry* Bloomington Indiana UP, 2000 vii, 270 p Cloth $44 95, paper $18 95

KEIL, FRANK C , and ROBERT A WILSON, eds *Explanation and Cognition* Cambridge MIT, 2000 x, 396 p Cloth $40 00

KENNY, ANTHONY, ed *The Oxford History of Western Philosophy* New York Oxford, 2001 vi, 450 p Paper $15 95

KRAUS, JODY S , and STEVEN D WALT eds *The Jurisprudential Foundations of Corporate and Commercial Law* New York Cambridge, 2000 245 p Cloth $54 95

KRZYSZTOF, ZIAREK, and SEAMUS DEANE, eds *Future Crossings* Evanston, IL Northwestern, 2000 309 p Cloth $79 95, paper $29 95

KUCZEWSKI, MARK G , and RONALD POLANSKY, eds *Bioethics Ancient Themes in Contemporary Issues* Cambridge MIT, 2000 xiii, 304 p Cloth $36 95

LARRIMORE, MARK, ed *The Problem of Evil* Malden, MA Blackwell, 2001 xxxii, 397 p Paper $29 95

LEITER, BRIAN, ed *Objectivity in Law and Morals* New York Cambridge, 2001 xi, 354 p Cloth $64 95

MOSER, PAUL K , and THOMAS L CARSON, eds *Moral Relativism A Reader* New York Oxford, 2001 ix, 337 p Cloth $45 00

NADLER, STEVEN, ed *The Cambridge Companion to Malebranche* New York Cambridge, 2000 xi, 319 p Cloth $54 95, paper $19 95

NAGL-DOCEKAL, HERTA, and CORNELIA KLINGER, eds *Continental Philosophy in Feminist Perspective Re-reading the Canon in German* University Park Penn State UP, 2000 vi, 354 p Paper $22 50

PADE, MARIANNE, ed *Renaissance Readings of the* Corpus Aristotelicum Copenhagen Museum Tusculanum, 2001 261 p Paper $34 00

PEIRCE, CHARLES S *Writings of Charles S Peirce A Chronological Edition, Volume VI 1886-1890* THE PEIRCE EDITION PROJECT, eds Bloomington Indiana UP, 2000 lxxxiv, 697 p Cloth $49 95

PETERS, MICHAEL, JAMES MARSHALL, and PAUL SMEYERS *Nietzsche's Legacy for Education Past and Present Values* Westport, CT Bergin and Garvey, 2001 xxx, 229 p Cloth $65 00

PETTEGREW, JOHN, ed *A Pragmatist's Progress? Richard Rorty and American Intellectual History* Lanham, MD Rowman and Littlefield, 2000 vii, 222 p Cloth $60 00, paper $24 95

PRESS, GERALD A , ed *Who Speaks for Plato? Studies in Platonic Anonymity* Lanham, MD Rowman and Littlefield, 2000 vi, 245 p Cloth $63 00, paper $23 95

RAPHAEL, FREDERIC, and RAY MONK, eds *The Great Philosophers* New York Routledge, 2000 469 p Cloth $35 00

REMPEL, RICHARD A , BERYL HASLAM, ANDREW BONE, and ALBERT C LEWIS *The Collected Papers of Bertrand Russell, Volume XV Uncertain Paths to Freedom Russia and China, 1919-22* New York Routledge, 2000 ciii, 684 p Cloth $200 00

ROGERS, G A J , and TOM SORELL *Hobbes and History* New York Routledge, 2000 xvi, 192 p Cloth $90 00

ROSS, DON, ANDREW BROOK, and DAVID THOMPSON, eds *Dennett's Philosophy A Comprehensive Assessment* Cambridge MIT, 2000 x, 395 p Cloth $65 00, paper $27 95

SCHACHT, RICHARD, ed *Nietzsche's Postmoralism Essays on Nietzsche's Prelude to Philosophy's Future* New York Cambridge, 2000 xiv, 264 p Cloth $54 95

SCHIEBINGER, LONDA, ed *Feminism and the Body* New York Oxford, 2000 ix, 500 p Cloth $24 95

SISTARE, CHRISTINE, LARRY MAY, and LESLIE FRANCIS, eds *Groups and Group Rights* Lawrence UP of Kansas, 2001 ix, 307 p Cloth $40 00, paper $17 95

SKRUPSKELIS, IGNAS K and ELIZABETH M BERKELEY, eds *The Correspondence of William James, Volume 8 1895 - June 1899* Charlottesville UP of Virginia, 2000 lxiii, 762 p Cloth $80 00

SMITH, NICHOLAS D , and PAUL B WOODRUFF, eds *Reason and Religion in Socratic Philosophy* New York Oxford, 2000 xiv, 226 p Cloth $45 00

SPERBER, DAN, ed *Metarepresentations A Multidisciplinary Perspective* New York Oxford, 2000 448 p Cloth $65 00, paper $35 00

STONE, M F W , and JONATHAN WOLFF, eds *The Proper Ambition of Science* New York Routledge, 2000 viii, 231 p Cloth $75 00

THOMPSON, PAUL B , and THOMAS C HILDE, eds *The Agrarian Roots of Pragmatism* Nashville Vanderbilt, 2000 ix, 342 p Cloth $39 95

VILLA, DANA, ed *The Cambridge Companion to Hannah Arendt* New York Cambridge, 2000 xvi, 304 p Cloth $54 95, paper $19 95

WRIGHT, JOHN P , and PAUL POTTER, eds *Psyche and Soma Physicians and Metaphysicians on the Mind-Body Problem from Antiquity to Enlightenment* New York Oxford, 2000 xii, 298 p Cloth $72 00

NEW BOOKS TRANSLATIONS

AQUINAS, THOMAS *Treatise on Law* RICHARD J REGAN, trans Indianapolis Hackett, 2000 xxvi, 102 p Cloth $19 95, paper $5 95

ARISTOTLE *Nicomachean Ethics* ROGER CRISP, ed and trans New York Cambridge, 2000 xlii, 213 p Cloth $49 95, paper $12 95, *Metaphysics Book B and Book K, 1-2* ARTHUR MADIGAN, SJ, trans New York Oxford, 1999 xl, 185 p Cloth $60 00, paper $24 95

BAYLE, PIERRE *Various Thoughts on the Occasion of a Comet* ROBERT C BARTLETT, trans Albany SUNY, 2000 xlviii, 332 p Paper $26 95

BLANCHOT, MAURICE, and JACQUES DERRIDA *The Instant of My Death* and *Demeure Fiction and Testimony* ELIZABETH ROTTENBERG, trans Stanford University Press, 2000 114 p Cloth $29 50, paper $12 95

BLOCH, ERNST *The Spirit of Utopia* ANTHONY A NASSAR, trans Stanford University Press, 2000 317 p Cloth $55 00, paper $19 95

BURLEY, WALTER *On the Purity of the Art of Logic The Shorter and the Longer Treatises* PAUL VINCENT SPADE, trans New Haven Yale, 2000 xxv, 323 p Cloth $60 00

CAVARERO, ADRIANA *Relating Narratives Storytelling and Selfhood* PAUL A KOTTMAN, trans New York Routledge, 2000 xxxii, 146 p Cloth $75 00, paper $22 95

CHANGEUX, JEAN-PIERRE, and PAUL RICOEUR *What Makes Us Think? A Neuroscientist and a Philosopher Argue about Ethics, Human Nature, and the Brain* M B DE BEVOISE trans Princeton University Press, 2000 336 p Cloth $29 95

DEBRAY, REGIS *Transmitting Culture* ERIC RAUTH, trans New York Columbia, 2000 ix, 157 p Cloth $29 50

DERRIDA, JACQUES, and ANNE DUFOURMANTELLE *Of Hospitality* RACHEL BOWLBY, trans Stanford University Press, 2000 x, 160 p Cloth $39 50, paper $14 95

FICHTE, J G *Foundations of Natural Right According to the Principles of the Wissenschaftslehre* FREDERICK NEUHOUSER, ed , MICHAEL BAUR, trans New York Cambridge, 2000 xxxv, 338 p Cloth $64 95, paper $22 95

HABERMAS, JÜRGEN *On the Pragmatics of Social Interaction Preliminary Studies in the Theory of Communicative Action* BARBARA FULTNER, trans Cambridge MIT, 2001 xxiv, 192 p Cloth $24 95

HONNETH, AXEL *Suffering from Indeterminacy An Attempt at a Reactualization of Hegel's* Philosophy of Right *Two Lectures* JACK BEN-LEVI, trans Amsterdam Van Gorcum, 2000 60 p Paper

HÖSLE, NORA K , and VITTORIO HÖSLE *The Dead Philosophers' Café An Exchange of Letters for Children and Adults* STEVEN RENDALL, trans Notre Dame University Press, 2000 vii, 166 p Cloth $19 95

JANICAUD, DOMINIQUE *The Theological Turn of French Phenomenology* VARIOUS TRANSLATORS New York Fordham, 2000 245 p Cloth $35 00, paper $20 00

KANT, IMMANUEL *Critique of the Power of Judgment* PAUL GUYER, ed , GUYER AND ERIC MATTHEWS, trans New York Cambridge, 2000 lii, 423 p Cloth $64 95

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO. 5, MAY 2001

THE MOST FAMOUS EQUATION*

A typical undergraduate physics textbook, written by Joseph W. Kane and Morton M. Sternheim,[1] says:

> The equation $E = mc^2$ is perhaps the most famous equation of twentieth-century physics. It is a statement that mass and energy are two forms of the same thing, and that one can be converted into the other (*ibid.*, p. 493).

With the first of these sentences, concerning the fame of Albert Einstein's equation, it is difficult to disagree. But the second sentence, which characterizes mass and energy as interconvertible and "two forms of the same thing," suggests that some care will need to be exercised in interpreting the famous equation. Indeed, it is difficult to find a scientific equation whose ontological implications have been misunderstood so widely and in so many ways. I shall begin by looking at some of the issues that arise when interpreting the famous equation. Along the way, I shall examine briefly some of the ways in which one could become confused by the equation's common interpretations—not only in ordinary physics texts, but also by some very distinguished physicists and philosophers. I shall then try to set the record straight.[2]

I. SOME ISSUES

Let us begin to lay out some of these interpretive issues by taking seriously the textbook's remark that mass "can be converted into"

---

* My thanks to Arthur Fine for comments on the penultimate draft of this essay, and to Alexander Rosenberg for his encouragement.

[1] *Physics* (New York: Wiley, 1978).

[2] To keep things simple and focused, I shall restrict myself throughout to interpretation of the special theory of relativity, leaving aside considerations arising from the general theory, quantum mechanics, and so forth.

0022-362X/01/9805/219–38

energy and vice versa Strictly speaking, this does not mean merely that when mass m disappears, a quantity E of energy takes its place, where $E = mc^2$ To say that some mass is "converted" into energy is to say that the mass *turns into* the energy that replaces it—that there is then a certain parcel of energy that used to be the mass

For example, after a radioactive nucleus decays, there is often said to be a "mass defect" the sum of the masses of the daughter bodies is less (by Δm) than the mass of the original nucleus Some (Δm) of the original mass is said to have been "converted" into the kinetic energy E of the daughter bodies, where $E = (\Delta m)\, c^2$ Since c (light's speed in a vacuum) is so large, a very small mass can be "turned into" a great deal of energy For instance, when a tritium nucleus (one proton, two neutrons) decays into a helium-3 nucleus (two protons, one neutron) along with an electron and an antineutrino, the tritium's mass exceeds the sum of the products' masses by a small quantity that is "equivalent" to about 0 0186 million electron volts of energy Has mass turned into energy, or merely disappeared and been replaced by an "equivalent" quantity of energy?

In classical physics, at least, a body's mass is often characterized as its "quantity of matter "[3] Mass, on this view, is a property, whereas matter is a stuff, a substance If mass is the quantity of matter, then perhaps the physics text should have said that *matter* is converted into energy Among the physical processes called to mind by this interpretation are a caterpillar being transformed into a butterfly and liquid water turning into steam

Energy, on this interpretation of the famous equation, would have to be like matter in being a stuff Classical physics sometimes seems to treat energy in that way Occasionally, energy conservation is stated as the principle that energy is "neither created nor destroyed" rather than as the principle that an isolated system's total energy remains constant A parcel of energy is sometimes described as being converted from one form into another or as moving from one place to another—for example as originally being stored as chemical energy in a battery, then passing through the electric and magnetic fields, eventually passing to a resistor (perhaps a flywheel, which possesses the parcel as kinetic energy), and finally (let us suppose) passing via friction to the surrounding air molecules as heat energy (that is, as the air molecules' kinetic energies of random motion).

---

[3] "*Quantity of matter is a measure of matter that arises from its density and volume jointly* I mean this quantity whenever I use the term 'body' or 'mass' in the following pages"—Definition 1 in Isaac Newton, *The* Principia *Mathematical Principles of Natural Philosophy*, I B Cohen and A Whitman, trans (Berkeley California UP, 1999), pp 403-04

This view of energy gained support in the late nineteenth century. Like other forms of potential energy, electric or magnetic potential energy had originally been considered as having an arbitrary zero level and as being possessed by a system as a whole rather than as having some definite distribution among the system's constituents. But with the rise of James Clerk Maxwell's electromagnetic theory, electric and magnetic potential energy were reinterpreted as electromagnetic field energy. Accordingly, a field of zero strength was taken to possess zero energy, and Maxwell's electromagnetic theory determined a field's energy density in space. All of this supported the view of energy as a substance that can flow, such motion requires that the energy at one place and time be the self-same thing as the energy at a certain other place and time.

On the other hand, many notable followers of Maxwell (such as Heinrich Hertz and Oliver Heaviside[4]) had grave doubts regarding what Heaviside called the "thingness" of energy. To mention just one source of disquiet: Maxwell's equations fail to fix uniquely the vector describing the flow of energy (specifically, the energy-flux density) through the electromagnetic field. The flux of the Poynting vector over an entire surface enclosing a volume is ordinarily interpreted as giving the rate at which electromagnetic-field energy is leaving the enclosed volume. But numerically the same result is given by the flux of the Poynting vector plus any vector field possessing zero divergence. While these two formulas for the energy-flux density agree on the net energy flow into or out of a given volume, such as the volume enclosed by a given spherical surface, the two formulas may disagree regarding the energy flow across a given open surface, such as the spherical surface's top hemisphere. (For instance, if the divergence-free vector field is characterized at every point by the same upward-pointing vector, then it portrays some additional energy as flowing into the volume through the bottom surface, as well as an equal additional energy flow out of the volume across its top surface. These two flows together have no net effect on the total enclosed energy.) The energy flow across an open surface is left entirely undetermined by classical electromagnetic theory, it might be termed "superfluous theoretical structure." Perhaps, then, there is no fact of the matter regarding the energy flow across an open surface—because energy is not a stuff (any more than velocity is)[5] Likewise, perhaps a body's mass is not the quantity of its matter, but rather merely a quantitative property determining the body's resistance to being accelerated by a force (This was Maxwell's[6] own understanding of mass) On this interpretation of mass and energy, we could not join Elie Zahar[7] in interpreting Einstein's famous equation as characterizing the underlying identity of what ordinarily appear to be two distinct substances

---

[4] *Electrical Papers, Volume 2* (Boston: Copley, 1925), p. 527.

> According to the accepted mechanistic view, ordinary "ponderable" matter, i e , the extended hard "stuff" with which we are all familiar, constitutes the most fundamental layer of physical reality [M]atter and energy were both conserved, but each separately from the other Einstein showed that the two levels could be regarded as identical The stuff which appears to the senses as hard extended substance and the quantity of energy which characterises a process are in fact one and the same thing (*ibid*, pp 261-62)

If energy and mass do not indicate quantities of stuffs the parcels of which have continuing identities over time, stuffs that can be transformed into one another in something akin to the sense in which a caterpillar turns into a butterfly, then it is difficult to see how there could be any reasonably strict sense in which mass *turns into* energy

Let us turn to another interpretive issue Our physics text said that $E = mc^2$ "is a statement that" mass can be converted into energy and vice versa But whether there happen to be, or even could physically possibly be, any processes that make mass disappear, putting energy in its place, is a different question from what the "rate of exchange" between mass and energy would be in such processes For instance, that certain nuclei are unstable, and that some mass is replaced by an "equivalent" quantity of energy when those nuclei decay, is a result of other laws, not just $E = mc^2$

That $E = mc^2$ tells us nothing about whether any physical processes allow energy and mass to be interconverted is exploited by Margaret Morrison[8] in a recent passage interpreting our favorite equation

> Does [$E = mc^2$] mean that mass and energy are the same thing? [I]n order for this relation to imply a complete identity, it seems that a total

---

[5] I examine this issue and a host of related problems in my *Locality, Fields, Energy, and Mass An Introduction to the Philosophy of Physics* (Malden, MA Blackwell, forthcoming), chapter 5

[6] *The Scientific Letters and Papers of James Clerk Maxwell, Volume 2,* P M Harman, ed (New York Cambridge, 1995), pp 396, 811, and Harman, "Edinburgh Philosophy and Cambridge Physics," in Harman, ed , *Wranglers and Physicists* (Manchester University Press, 1985), pp 204-24, especially pp 222-23

[7] *Einstein's Revolution A Study in Heuristic* (La Salle, IL Open Court, 1989)

[8] *Unifying Scientific Theories* (New York Cambridge, 2000)

> conversion of mass into energy should be possible However, such a conversion is not possible, because of the conservation of baryon number, which says that the total number of these particles cannot change, they can be converted into other baryons, but cannot disappear entirely Baryons include, among other things, protons and neutrons, which make up most of the mass of ordinary matter, and therefore most of the mass of ordinary matter is not available for conversion into energy So we are left in the rather odd position of asserting an equivalence between mass and energy, but not an identity that would allow one to be fully transformed into the other (*ibid*, p 182)

These remarks correctly recognize that $E = mc^2$ is logically consistent with limitations imposed by other laws on the processes that allow mass to be converted into energy But these remarks may mislead us regarding what these limitations actually are Baryon number conservation does not prevent baryons (such as protons and neutrons) from being "fully transformed" into energy (or vice versa) A proton has baryon number $+1$ and an antiproton has baryon number -1, so their mutual annihilation (with the creation of photons, quanta of light possessing energy and having baryon number 0) does not violate baryon number conservation Of course, antiprotons are not "ordinary matter," and so Morrison is correct insofar as she is suggesting that without the assistance of extraordinary matter, most of the mass of ordinary matter is not available for conversion into energy

Nevertheless, I do not see why Morrison insists that for mass and energy to be the same thing, a total conversion of mass into energy must be physically possible Even if mass and energy were essentially the same (like water and steam), it could be a natural law that 90% is the maximum fraction of the initial mass that a process can turn into energy Or it could simply be that there is no process, permitted by the laws of nature, that turns mass entirely into energy, just as it could be that there is no process, permitted by the laws of nature, that enables diamond to be turned into graphite (though both are forms of carbon)

We have now arrived at the central claim of the textbook passage that according to $E = mc^2$, "mass and energy are two forms of the same thing " Some authors try to slide from the interconvertibility of mass and energy to their identity Here, for instance, is Max Jammer[9]

> As we have seen, the conversion of an electron-positron pair, for example, into gamma radiation or its mirror phenomenon is an incontestable experimental confirmation of the assertion of the theory of relativity that

---

[9] *Concepts of Mass* (Cambridge Harvard, 1961)

> mass and energy are mutually and completely interconvertible This state of affairs raises the following questions Are not the two entities which are interchangeable essentially the same? Is not what is generally spoken of as an equivalence in reality an identity? Are therefore not "mass" and "energy" merely synonyms for the same physical reality, which may perhaps be termed "massergy" (*ibid*, p 184)?

Jammer clearly intends the answer to be "yes," but from the fact that one thing can be converted into the other, it simply does not follow that the two are actually different forms of the same thing James Prescott Joule, for instance, famously suggested that all forms of energy (heat, kinetic energy, gravitational potential energy, chemical bond energy, electrical energy, and so on) are interconvertible, but he did not go so far as to hold that they are all different forms of the same thing In 1843, he showed experimentally that (what he called) the "mechanical equivalent of heat" is (in modern units) 4 18 Joules of energy per calorie of heat In other words, Joule demonstrated that a 1 kg mass falling through a distance of 42 67 cm contributes exactly enough mechanical energy (in doing work by, say, turning a stirring paddle immersed in water) to add 1 calorie of heat to the water But he saw this "equivalence" as involving how much of one thing replaces a given quantity of the other—their fixed "rate of exchange " He did not understand the mechanical equivalent of heat as we do today as a conversion factor, that is to say, a way to change our units for measuring the same thing, because heat is just a particular form of energy [10] Yet this is how "mass-energy equivalence" is often understood In the words of a fine relativity text by Edwin Taylor and John Archibald Wheeler[11]

> [J]oules and kilograms are two units—different only because of historical accident—for one and the same kind of quantity, mass-energy The conversion factor c^2, like the factor of conversion from miles to feet, can today be counted, if one wishes, as a detail of convention, rather than as a deep new principle (*ibid*, p 137)

As Einstein[12] himself writes

---

[10] See Yuhuda Elkana, *The Discovery of the Conservation of Energy* (Cambridge Harvard, 1974) Einstein writes "[E]ach absorption or release of energy brings about, respectively, an increase or decrease of the mass of the body involved Energy and mass appear as equivalent quantities in the same way that heat and mechanical energy do"—"On the Development of Our Views Concerning the Nature and Constitution of Radiation," in *The Collected Papers of Albert Einstein, Volume 2*, A Beck, trans , pp 379-94, on p 386

[11] *Spacetime Physics* (San Francisco Freeman, 1966)

[12] *The Meaning of Relativity* (Princeton University Press, 1953)

> Mass and energy are therefore essentially alike, they are only different expressions for the same thing (*ibid*, p 45)

But how is mass-energy equivalence in this ontological sense (not merely in terms of a "rate of exchange") to be understood?

One problem arises from the fact that under a standard interpretation, relativity theory denies the objective reality of various properties that we ordinarily assign to material bodies (such as their length and velocity) and to events (such as their separation in space and their separation in time) Each of these quantities is *frame dependent*; none is "Lorentz invariant"—that is, the same in every inertial frame of reference Only what is the same in every inertial frame is a genuine feature of reality The value that any frame-dependent quantity assumes in a given inertial frame reflects not just reality, but also that reference frame's own particular perspective The Lorentz invariant quantities are exactly those which depend only on how the universe really is, uncontaminated by any contribution from us in describing the universe

Consider a body's mass m (Here I mean what is sometimes called the body's "rest mass"—though, as I shall explain, I prefer to call it simply the body's "mass") This quantity is Lorentz invariant, and hence objectively real But the body's energy is not, since its energy depends on its speed, and its speed v is plainly frame dependent[13] In particular, relativity says that a body's energy E is $m\gamma c^2$, where $\gamma = 1/\surd[1 - (v^2/c^2)]$ Indeed, Heaviside made a similar point in a classical context, where a body's kinetic energy is $(1/2)mv^2$ One reason that Heaviside doubted energy's substantiality is that a body's energy depends on its velocity and so is frame dependent, and Heaviside believed in what he called "the relativity of motion" that there are in reality no absolute velocities since only what is the same in every inertial frame can be a genuine feature of reality (*op cit*, pp 525-26, 521)[14] Likewise, the path taken by energy flowing in a system, being frame dependent, lacks objective reality

A body's energy E and momentum p (a vector—I suppress any vector notation—the magnitude of which, in relativity theory, is $m\gamma v$), *taken together*, form an invariant, mc

$$\surd[(E/c)^2 - p^2] = \surd[(m\gamma c)^2 - (m\gamma v)^2] = m\gamma\surd[c^2 - v^2]$$

$$= m(1/\surd[1 - (v^2/c^2)])c\surd[1 - (v^2/c^2)] = mc$$

---

[13] At this point, you might wonder how E can possibly equal mc^2 when E is frame dependent but m and c are the same in every inertial frame Stay tuned

[14] See also Heaviside, *Electromagnetic Theory, Volume 1* (London Benn, 1922), p 75

To appreciate this invariant, let us use space-time geometry At a given moment, a body is associated with an energy-momentum 4-vector P The lengths of P's three spatial components (along the x, y, and z axes) are given by the components p_x, p_y, and p_z of the body's momentum, which equal $m\gamma v_x$, $m\gamma v_y$, and $m\gamma v_z$ (where v_x, v_y, and v_z are, respectively, the x, y, and z components of the body's velocity) In other words, the body's momentum p is the 3-vector formed by P's spatial components The temporal component of P is $m\gamma c$, which is E/c In Euclidean geometry, the square of the hypotenuse's length is, of course, the sum of the squares of the other two sides' lengths But in space-time geometry, the square of a 4-vector's length is the square of its temporal component's length *minus* the square of the length of the 3-vector formed by its spatial components Accordingly, the length of the 4-vector P is $\surd[(E/c)^2 - p^2]$, which (as we just saw) equals the invariant mc In short, P's length is an invariant quantity, though its decomposition into energy and momentum components is frame dependent, just as the electromagnetic field appears as different combinations of electric and magnetic fields from the standpoints of different inertial frames Relativity thus unifies energy and momentum in exactly the same sense as it unifies the electric and magnetic fields[15]

Classically, a body's momentum $p = mv$, whereas relativistically, $p = m\gamma v$ By calling a body's $m\gamma$ its "relativistic mass" M, momentum in relativity remains mass times velocity, and so an analogy can be drawn between relativistic and classical physics Whereas m is invariant, M is not, since γ is obviously frame dependent in being a function of v One version of our favorite equation, $E = Mc^2$, then follows from $E = m\gamma c^2$ The celebrated "$E = mc^2$," featuring the body's (invariant) mass m, holds only when the body is at rest (that is, when its $p = 0$)[16] since $E = mc^2$ then follows from the more general

$$\surd[(E/c)^2 - p^2] = mc$$

When the body is at rest ($v = 0$), its $\gamma = 1$, and so its "relativistic mass" $m\gamma$ equals its mass m That is why "m" is sometimes called the body's

---

[15] That p is conserved in every inertial frame entails that E is so conserved as well, since p in one frame is a function of p and E in another But this entailment alone does not suffice to unify momentum and energy—does not reveal them to be nothing more than aspects of the same single real thing—any more than momentum, energy, and mass are unified in classical physics by the fact that momentum conservation follows from energy conservation and mass conservation in every inertial frame

[16] Thus the question posed in footnote 13 is answered $E = mc^2$ holds only for E in the $p = 0$ frame

"rest mass" its "relativistic mass" when it is at rest. Because the so-called "relativistic mass" is not an invariant quantity (and the term 'rest mass' refers back to 'relativistic mass'), the best thing to do in order to avoid confusing frame-dependent quantities with invariant ones is just to avoid the terms 'relativistic mass' and 'rest mass', and instead to stick solely with 'mass' for the invariant quantity symbolized m.[17]

Now we can ask a question regarding the proper interpretation of $E = mc^2$. Given that mass is a real property (since it is Lorentz invariant) whereas energy is not, how can mass and energy be the same thing ("mass-energy") *in the same sense* that distance in miles measures the same real thing as distance in feet? Moreover, in what sense can mass be *converted* into energy (as when a tritium nucleus decays) when mass and energy are not on a par in terms of their reality? We shall look into this.

On the basis of $E = mc^2$ (when $p = 0$), Einstein[18] says that "[t]he mass of a body is a measure of its energy content" (in the $p = 0$ frame). But beware: this remark may mislead us by suggesting that energy content is a basic, objective fact whereas mass is merely some sort of indicator, symptom, or manifestation of energy. In fact, however, mass rather than energy is the Lorentz invariant, and hence real quantity.[19] Accordingly, I cannot agree with remarks like Bertrand Russell's[20]

> [W]e must abandon attempts to say what energy is. We must say simply there is something quantitative, to which we give the name "energy"; this something is very unevenly distributed in space; there are small regions in which there is a great deal of it, which we call "atoms", and are those in which, according to older conceptions, there was matter. ... Mass is only a form of energy. ... It is energy, not matter, that is fundamental in physics (*ibid.*, pp. 27, 291).

---

[17] That Einstein did not regard "rest mass" and "relativistic mass" as standing on the same ontological footing is evident from his many remarks urging others to introduce as "mass" only a quantity independent of motion. See Arthur Fine, "Appendix" to John Earman, "Against Indeterminacy," this JOURNAL, LXXIV, 9 (September 1977): 535-38, p. 538; and Lev Okun, "The Concept of Mass," *Physics Today*, XLII, 6 (June 1989): 31-36.

[18] "Does the Inertia of a Body Depend Upon Its Energy Content?" in *The Collected Papers of Albert Einstein, Volume 2*, pp. 172-74, see p. 174. See also his *Relativity: The Special and the General Theory* (New York: Crown, 1931), pp. 55-56.

[19] Of course, there is an objective fact regarding energy content *in the $p = 0$ frame*, but this is a cheap sort of objectivity: it holds in all frames because it refers to a specific frame.

[20] *Human Knowledge: Its Scope and Limits* (New York: Simon and Schuster, 1948).

How can mass really be nothing but concentrated energy[21] that moves through space, "a form of bound energy,"[22] when mass rather than energy is the invariant quantity? Because energy is not Lorentz invariant, it seems difficult to endorse these remarks from Werner Heisenberg[23]:

> Since mass and energy are, according to the theory of relativity, essentially the same concepts, we may say that all elementary particles consist of energy. This could be interpreted as defining energy as the primary substance of the world (*ibid.*, p. 67).

This kind of talk filters down to monographs for guiding physics teachers, such as one by Robert M. Cotts and Robert W. Detenbeck,[24] which says, "Energy appears in still another wonderful disguise as mass itself" (*ibid.*, p. 66).

Here is a final issue regarding the ontological significance of Einstein's famous equation. In any inertial frame, a closed system's total energy and total momentum (as defined relativistically) are each "conserved." That is, in any inertial frame, the system's total energy does not change as time passes, and likewise for its total momentum. (But these totals differ in different inertial frames; total energy and total momentum are frame dependent and therefore not real.) Since E and p are conserved and we just saw an equation by which m is determined by E and p, a closed system's total m must be conserved in any inertial frame. But we saw earlier that when mass "turns into" energy (as in the decay of a tritium nucleus), there is a "mass defect" and so, apparently, mass fails to be conserved. (In these cases, textbooks commonly refer to some "missing mass" Δm.) We must look into how mass conservation can be reconciled with "mass defects."

II. WHAT IS MASS?

In classical physics, a body's mass is often interpreted as the amount of some "stuff" (matter) of which the body is made. In relativity, however, this interpretation cannot be correct. That is because properties that represent the quantity of some "substance" must obey the following principle: the total quantity of some sort of stuff in a system is the sum of the various quantities of that stuff belonging to the system's parts (where those parts are finite in number, nonoverlapping, and together exhaust the system).[25] For example, since the density of a whole is not the sum of the densities of its parts, density does not measure the amount of some stuff, and similarly for temperature and velocity. On the other hand, mass in classical physics is "additive" in this way. But in relativity, mass is not additive.

---

[21] Hermann Weyl, *Space, Time, Matter* (New York: Dover, 1922), p. 200.

[22] Hans Pagels, *The Cosmic Code* (New York: Simon and Schuster, 1982), p. 37.

[23] *Physics and Philosophy* (London: George Allen and Unwin, 1958). The difficulties for understanding energy as a locally well-defined, conserved quantity become even more substantial as we pass to the general theory of relativity. See Carl Hoefer, "Energy Conservation in GTR," *Studies in History and Philosophy of Modern Physics* 31B, 2 (June 2000): 187-99.

[24] *Matter in Motion* (Seattle: Washington UP, 1966).

To see this, let us work in an inertial frame in which $p = 0$. (There always exists such a frame, and since mass is invariant, the result in that frame is applicable to any other inertial frame.) Then (as I have already mentioned)

$$E = mc^2$$

Consider a system of finitely many constituents, where each exerts on the others only negligible forces when they are not in contact—such as a gas consisting of many molecules.[26] Since the system's total energy is the sum of the energies E_1, E_2, … of its constituents, we have

$$m = (1/c^2)[E_1 + E_2 + \ldots]$$

Recall that for any constituent (say, the ith one),

$$E_i = m_i\gamma_i c^2 = m_i c^2/\sqrt{[1 - (v_i^2/c^2)]}$$

Now for any x where $x^2 < 1$,

$$1/\sqrt{[1 - x]} = 1 + (1/2)x + [(1\cdot 3)/(2\cdot 4)]x^2 + [(1\cdot 3\cdot 5)/(2\cdot 4\cdot 6)]x^3 + \ldots$$

Let x be (v_i^2/c^2). If v_i is low compared to c, then the terms with x^2 and higher powers of x are negligible, and so we can use the approximation

$$E_i \approx m_i c^2 + (1/2)m_i v_i^2$$

Notice that this is just $m_i c^2$ plus the ith constituent's classical kinetic energy. Substituting for E_1, E_2, …,

$$m \approx (1/c^2)[m_1 c^2 + (1/2)m_1 v_1^2 + m_2 c^2 + (1/2)m_2 v_2^2 + \ldots]$$

and so

---

[25] See Norman Robert Campbell, *Foundations of Science* (New York: Dover, 1957), pp. 282-83, 286 [originally *Physics: The Elements* (New York: Cambridge, 1920)]. For related discussion, see Helen Morris Cartwright, "Amounts and Measures of Amount," *Noûs*, ix, 2 (May 1975): 143-64, pp. 147-49.

[26] This restriction means that there is no potential energy, there are no fields, and so forth; in classical terms, the only energy is kinetic energy. I relax this restriction in chapter 8, section 5 of my *Locality, Fields, Energy, and Mass* and go on to examine the mass possessed by the electromagnetic field.

$$m \approx [m_1 + m_2 + \quad] + (1/c^2)[(1/2)m_1v_1^2 + (1/2)m_2v_2^2 + \quad]$$

Hence, the system's mass m exceeds the sum of its constituents' masses (They differ by an amount that reflects the constituents' kinetic energies in the frame where the system has zero total momentum This will shortly be important) So we cannot think of a body's mass as the amount of matter that forms it, where matter is a "stuff "

Note that the system's mass *is* the sum of its constituents' *relativistic* masses $m_i\gamma_i$ in the frame where $p = 0$ That is because in that frame,

$$m = (1/c^2)[E_1 + E_2 + \quad] = (1/c^2)[m_1\gamma_1c^2 + m_2\gamma_2c^2 + \quad]$$

$$= m_1\gamma_1 + m_2\gamma_2 +$$

But since a body's relativistic mass is frame dependent, it is not the body's quantity of any sort of real "stuff " So the fact that the system's (invariant) mass is the sum of its constituents' relativistic masses in the $p = 0$ frame does not show that the system's mass is its total quantity of some sort of "stuff" of which it is made

Bearing in mind that mass is not additive, consider the following argument, some version of which appears in most relativity texts Suppose two bodies, each of mass m_0, crash into each other and stick together, forming one body at rest of mass m_1 In the words of one fine text by Robert Resnick and David Halliday[27]

> [Although] the total energy E of the system of particles will be conserved [and] relativistic mass is conserved in this collision, *rest mass* [what I am calling simply "mass"] is not, m_1, the rest mass after the collision, is greater than $2m_0$, the rest mass before the collision After the collision, no kinetic energy remains In place of the "lost" kinetic energy, there appears internal (thermal) energy, recognizable by the rise in temperature of the colliding particles [By $E = mc^2$,] (increase in thermal energy) = (increase in rest mass)(c^2) Thus we see that the decrease in kinetic energy for this isolated system is balanced by a corresponding increase in mass energy (*ibid*, p 110)

So the text says that although the system is "isolated," mass is not conserved! This is not correct In this argument, the sum $2m_0$ of the masses before the collision is shown to be unequal to the mass m_1 of the body formed in the collision But $2m_0$ is not "the rest mass before the collision" (that is, the mass of the system before the collision) because *mass is not additive* The argument that mass is not conserved rests on the mistaken assumption of mass's additivity To say that energy is conserved whereas mass is not helps to make energy seem like some sort of real stuff, leading to remarks like Heisenberg's (already quoted). Of course, it is also simply confusing for the text I have quoted to say on one page that the isolated system's rest mass is not conserved and on the very next page that "if we consider a closed system then we may regard the rest mass of the body (or of the system) as constant."

---

[27] *Basic Concepts in Relativity and Early Quantum Theory* (New York Wiley, 1985, 2d ed) I have substituted 'm_1' for their symbol for the resultant body's mass

If a body's mass is not its total quantity of some sort of stuff of which it is made, what is a body's mass? A body's mass is the property it possesses which determines the acceleration it undergoes in response to a force $F = p'$ (as in classical physics), where in relativity, $p = m\gamma v$ and p' is the rate at which p is changing. When referring to a body's "mass," then, we must be thinking of that body as a thing that can feel a force and respond to it (by moving) *as a unit*. Nevertheless, no macroscopic body is elementary; any macroscopic body is also a system of bodies. Its motion, then, is nothing but the motions of its constituents, and these motions are determined by *their* masses and the forces that *they* feel. The remarkable fact is that the law of nature by which the constituents' motions are determined by their masses and the forces they feel is *the same as* the law by which the macroscopic body's motion is determined by its mass and the forces it feels: $F = p' = (m\gamma v)'$. In other words, the law "scales up."[28]

Let us see precisely why this fact is so remarkable. Macroscopic bodies are presumed to have elementary constituents (that is, constituents having no constituents themselves) and there is a law relating the force felt by an elementary constituent, the constituent's motion in response to that force, and a *single parameter* characterizing the constituent. This need not have been; we can imagine a universe in which a constituent's motion in response to forces is determined in different ways by several of its properties (such as its shape, volume, and chemical activity) rather than there being a single parameter (its "mass") that captures everything about the constituent affecting its motion in response to a force. But from the fact that such a parameter exists for elementary bodies, it does not follow that a single parameter characterizing the macroscopic body is all that is needed to allow its motion to be related to the force that it experiences. If a body's "mass" is defined as that single parameter, then even given the remarkable fact that the fundamental constituents of macroscopic bodies have masses, it is a remarkable fact that macroscopic bodies *have masses too*.

---

[28] A similar remark appears in Richard Feynman, Robert Leighton, and Matthew Sands, *The Feynman Lectures on Physics, Volume 1* (Reading, MA: Addison-Wesley, 1963), p. 19-2. For related discussion, see David Bohm, *The Special Theory of Relativity* (New York: Routledge, 1996), pp. 82-85, 110-18.

Furthermore, from the fact that macroscopic bodies have constituents with masses *and* that macroscopic bodies have masses too, it does *not* follow that the law by which the elementary constituents' motions are determined by their masses and the forces they feel is *the same as* the law by which the macroscopic body's motion is determined by its mass and the forces it feels. If a body's "mass" is defined as the quantity with which it is associated that plugs into the particular law relating an *elementary* body's motion to the forces it feels, then once again, it is a remarkable fact that macroscopic bodies *have masses too*.

In short, neither of these conceptions of what mass *is* requires that macroscopic bodies have masses, even considering that their elementary constituents do. Furthermore, given either of these conceptions of mass *and* that a macroscopic body has a mass, there is nothing inevitable about its mass being the sum of its elementary constituents' masses (as is the case classically but not relativistically). So these conceptions of mass should inoculate us against the temptation to think that a body's "mass" *is defined as* the quantity of matter composing it.

It is certainly convenient for us that in order to predict a macroscopic body's motion in response to a force exerted upon it from outside it (an "external force"), we do not need to determine how each of its constituents responds to the various forces exerted upon it by other constituents ("internal forces"). We can simply think of the macroscopic body's mass m as if it were concentrated in a point particle located at the body's "center of mass" and then determine this particle's response to the external force. It is also convenient that the law governing this response is the same as the law governing the responses of the body's elementary constituents to the forces they feel. Otherwise, our investigations into the behavior of macroscopic bodies would not necessarily have left us well equipped to understand the behavior of their elementary constituents.

Let us see exactly why in classical physics, Newton's second law ($F = ma$, where F is a force felt by a body of mass m and a is the body's acceleration in response to that force) "scales up" in that it governs the motions not only of the elementary bodies, but also of the center of mass of a system of those bodies. To simplify the discussion, suppose there are only three elementary bodies, each exerting forces on the others. Let m_i be the ith body's mass, and let a_{ij} be the component of the ith body's acceleration that is caused by the force F_{ij} exerted by the jth body on the ith body. By Newton's second law,

$F_{ij} = m_i a_{ij}$ By Newton's third law ("Every action has an equal and opposite reaction"), $F_{ij} = -F_{ji}$, and so $m_i a_{ij} = -m_j a_{ji}$ Now consider bodies 2 and 3 as forming a single system of bodies Let us see why that system's motion is likewise governed by Newton's second law The system's center of mass is a kind of average of the positions of the system's constituents Each constituent's contribution to the average is proportional to its mass So the system's center of mass is a *weighted* average of its constituents' positions—weighted by their masses If x_i is the position of the ith body, then the position x of the center of mass of the system composed of bodies 2 and 3 is given by

$$x = [m_2 x_2 + m_3 x_3]/[m_2 + m_3]$$

For Newton's second law to govern this system's motion would be for the force exerted on the system (namely, the sum of the forces exerted on the system's constituents by body 1, the body outside the system $F_{21} + F_{31}$) to equal the system's mass ($m_2 + m_3$, in classical physics, mass is additive) multiplied by the acceleration a of its center of mass From the above equation for x, it follows that

$$a = [m_2 a_2 + m_3 a_3]/[m_2 + m_3]$$

So Newton's second law governs the system's motion if and only if

$$F_{21} + F_{31} = (m_2 + m_3)a = (m_2 + m_3)[m_2 a_2 + m_3 a_3]/[m_2 + m_3]$$
$$= m_2 a_2 + m_3 a_3$$

By Newton's second law applied to the constituents,

$$m_2 a_2 = F_{21} + F_{23}$$
$$m_3 a_3 = F_{31} + F_{32}$$

So Newton's second law governs the system's motion if and only if

$$F_{21} + F_{31} = F_{21} + F_{23} + F_{31} + F_{32}$$

But this is obviously true since (by Newton's third law)

$$F_{23} = -F_{32}$$

So the system consisting of bodies 2 and 3 behaves in relation to body 1 in just the way that one elementary body behaves in relation to another (By expanding the system to encompass more and more elementary bodies, we could work up to the conclusion that macroscopic bodies obey Newton's second law) Thus, it takes the cooperation of Newton's third law for Newton's second law to scale up Since either or both these laws could have been different, there is nothing logically mandatory about the fact that the law relating a body's motion to the forces it feels scales up.

It is likewise the case in relativity theory that a collection of bodies behaves as a single body with a mass m located at the collection's center of mass. According to relativity, the system's center of mass is the weighted average of its constituents' positions—weighted by their energies (or, what comes to the same thing, their relativistic masses):

$$x = [m_1\gamma_1 x_1 + m_2\gamma_2 x_2 + \ldots]/[m_1\gamma_1 + m_2\gamma_2 + \ldots]$$

(This formula is restricted to the case where the constituents exert forces on one another only when they are in contact—in other words, when no fields are involved. At the start of this section, I restricted myself to this case for the sake of simplicity.) For example, if the system is closed (that is, feels no external forces), then (since it behaves as a single body) its center of mass undergoes no acceleration. Notice that the weights in the weighted average are frame-dependent quantities (since the *i*th weight involves γ_i, which is a function of the *i*th body's velocity, which is obviously frame dependent). So the location of the system's center of mass is not Lorentz invariant.

Now we are in a position to appreciate the key point.

III. THE CONVERSION OF MASS TO ENERGY IS NOT A PHYSICAL PROCESS (OR VICE VERSA)

The frame in which the system's total $p = 0$ is the frame in which the system's center of mass is at rest. As we have already seen, the system's mass increases as the kinetic energies of its constituents increase in the $p = 0$ frame—that is, as its constituents move about more quickly relative to the system's center of mass. That is because in this frame, the system's

$$m = E/c^2 = (1/c^2)[E_1 + E_2 + \ldots]$$

and so any increase ΔE in the constituents' total kinetic energy contributes $\Delta E / c^2$ to the system's mass, though it has no effect on any constituent's mass. Imagine, for example, a ball of hot gas, its molecules whizzing around randomly. We add to the constituents' total kinetic energy by heating the gas. We thereby increase the gas's mass. In what sense is this the "conversion" of energy into mass?

Suppose that initially we think of the gas as a collection of molecules, each with its own mass. In other words, we treat each of these bodies as individually feeling forces and being accelerated by them, its acceleration depending on its mass. Accordingly, we characterize the heat as having boosted various molecules' kinetic energies, but not their masses Suppose we then change our perspective by considering the ball of gas as a single body that feels external forces and moves about as a unit From this perspective, the kinetic energy contributed by the heat becomes part of the gas's mass This "conversion" of energy into mass is not the transformation of one kind of stuff into another, since neither energy nor mass is a kind of stuff or measures the amount of some kind of stuff (energy because it is not Lorentz invariant, mass because it is not additive) But more importantly, we have just seen that this "conversion" of energy into mass is not any kind of real physical process taking place in nature *We* "converted" energy into mass simply by *changing our perspective* on the gas from treating it as many bodies to treating it as a single body

Let us see that again Suppose we begin by treating the gas as a single body The body is heated Heat energy flows into it and its mass increases by an equivalent amount It looks like energy is being converted into mass, fluid, gossamer energy has "solidified" or "congealed" into matter, "the extended hard 'stuff' with which we are all familiar" (in Zahar's words) I argue, however, this is not a real process, rather, it is just an artifact of the perspective we have adopted No such "conversion" occurs on a different perspective Let us begin the gas's story again and this time, let us treat the gas as many bodies We find no energy being transformed into matter as the heat is being added to the gas—so long as we continue to regard the gas as many bodies, none of those bodies increases its mass while the gas is heating up The heat energy goes into their kinetic energies relative to the gas's center of mass.

So on the first perspective, energy was converted into mass, whereas on the second, no such conversion occurred Furthermore, at whatever point in the story we choose to switch our perspective on the gas, we can make it appear that at that point, energy is being transformed into mass—even if at that point, heat is no longer being added to the gas! Again let us begin by treating the gas as many bodies, and let us maintain this perspective throughout the heating of the gas Suppose the heat source is then shut off Nothing more is now happening to the gas But we decide at this point to shift our perspective in telling the gas's story, henceforward we treat the gas as a single body In this shift, the gas molecules' total kinetic energy (in the $p = 0$ frame) contributes an equivalent amount to (and so "becomes part of") the gas's mass But obviously, no physical change accompanies this "transformation", again, nothing is happening to the gas This "conversion" of energy into mass is not a physical process Thus, whether and when a conversion of energy into matter occurs in the story of the gas depends on the perspective from which we elect to tell that story and any shifts of perspective we make in the course of telling it.

The distinction between forms of "internal" and "external" energy is often drawn by relativity texts.[29] But the point is seldom made that because this distinction is a scientific convention, so is the "conversion" of energy to mass or vice versa. As far as science is concerned, the line between bodies—between "internal" and "external"—is a convention (that is, not built into the universe, but rather drawn onto it by us) because of the remarkable fact that I emphasized in the previous section: that *any* system is characterized by a *mass* determining the way it responds (by moving as a unit) to the forces it feels as a whole (that is, to external forces). As Einstein[30] says: "Every system can be looked upon as a material point as long as we consider no processes other than changes in its translation velocity as a whole" (*ibid.*, p. 225).

A shift in "perspective" (*not* reference frame!), in what are being treated as single bodies, obviously takes place in the course of the textbook discussion (quoted earlier) of two bodies colliding and subsequently sticking together. The text treats the system as initially consisting of two bodies, each with its own mass. But the text regards the system after the collision as forming a single body. The "conversion" of energy into mass in this case is an artifact of this shift in perspective.

Sometimes, a shift in perspective is much subtler. Return to the tritium nucleus (one proton, two neutrons) that decays into a helium-3 nucleus (two protons, one neutron) and an electron and antineutrino that fly off at high speed. There is a "mass defect" in that the masses of a helium-3 nucleus, an electron, and an antineutrino add up to less than the mass of a tritium nucleus. The missing mass is said to have been "converted" to the kinetic energies of the resulting bodies. But this "conversion" of mass into energy is not real, it is an illusion produced by a subtle shift in our perspective. (The transformation of the tritium's neutron into a proton, an electron, and an antineutrino is, of course, a real occurrence.) We treated the system as initially forming a single body: a tritium nucleus. But we treated the system after the decay as consisting of three bodies, each with its own mass. The system's mass after the decay is the same as the system's mass before the decay. There is no "mass defect" here, mass is conserved. The "mass defect" appears to arise from the fact that the sum of the three masses after the decay is less than the system's mass before the decay (the difference reflecting the three bodies' kinetic energies in the $p = 0$ frame) But the sum of the three masses after the decay is less than the *system's* mass *after* the decay Mass is not additive, and our expectation that it is additive (arising because we expect it to measure the amount of some stuff) leads us to refer to a "mass defect"—to ask where the "missing mass" has gone and to conclude that it has turned into energy The "mass defect" results not from some physical transformation of matter-stuff into energy-stuff, but rather from our illicitly trying to view the system from two different "perspectives" at the same time

---

[29] See, for example, Resnick and Halliday, pp. 111-12.

[30] "Elementary Derivation of the Equivalence of Mass and Energy," *Bulletin of the American Mathematical Society*, XLI, 4 (April 1935): 223-30.

Thus, the conservation of mass does not conflict with the original tritium nucleus's mass exceeding the sum of the postdecay bodies' masses The "mass defect" is not real, it is produced not by the decay of the tritium nucleus, but by our treating the postdecay system as a collection of bodies even while treating the predecay system as a single body The fact that Δm of the system's initial mass "becomes" energy $(\Delta m) c^2$ when we think of the postdecay system as a collection of bodies, each with its own mass, does not mean that mass is really nothing but energy or that mass and energy are like distance in feet and in miles—different ways of measuring the same property Again, the "conversion" of mass into energy occurs because we have shifted our perspective, not because the nucleus has decayed Unlike the conversion of a caterpillar into a butterfly, the conversion of mass into energy (or vice versa) is not a physical process

A body's energy $m\gamma c^2$ is its kinetic energy plus mc^2 Since this latter energy depends only on its mass, we can think of its mass as "associated with" a certain quantity of energy a "mass energy" mc^2 Accordingly, it might sometimes help to think of relativity as having turned the classical law of energy conservation into the conservation of "mass-energy" by having added a term for the energy "frozen" as a body's mass In Einstein's[31] words.

> Before the advent of relativity, physics recognised two conservation laws of fundamental importance, namely, the law of conservation of energy and the law of conservation of mass, these two fundamental laws appeared to be quite independent of each other By means of the theory of relativity they have been united into one law (*ibid*, p 54)

But the fact that, in totaling a system's energy, we must include not only the more familiar forms of energy but also terms of the form mc^2

---

[31] *Relativity The Special and the General Theory*

does not mean that mass really *is* energy (or "massergy") or "that mass is now viewed as a form of energy"[32] like ice is a form of water Mass is a real property whereas energy is not Mass-energy is not some sort of real stuff that is neither created nor destroyed

Just as there is only a single object, the electromagnetic field, which in different inertial frames appears as different combinations of electric and magnetic fields, so there is only a single real property, the body's mass, which in different frames appears as different combinations of its energy and momentum This is the sense in which relativity theory unifies energy, momentum, and mass

MARC LANGE

University of Washington

---

[32] Marshall Spector, *Methodological Foundations of Relativistic Mechanics* (Notre Dame University Press, 1972), p 151

DOES PERCEPTION HAVE A NONCONCEPTUAL CONTENT?*

The question posed in my title is one that has been vigorously debated in philosophy for almost twenty years now In one form or another, the idea that perceptual experience has a content that is nonconceptual is found in the writings of, among others, José Bermúdez, Tim Crane, Fred Dretske, Gareth Evans, Susan Hurley, myself, and Michael Tye [1] The idea has been strongly opposed by John McDowell, Sonia Sedivy, and, most recently, by Bill Brewer [2] The question has generated so much discussion because the possibility—or otherwise—of nonconceptual content is inseparable from some fundamental issues about the individuation of conceptual content, about the nature of concept possession, about the nature of rationality, about the relation between animal and human perception, and even about our conception of objectivity—which all turn, in part, on the possibility of nonconceptual content in perception That my question is inseparable from these wider issues I shall try to make

---

* The present paper was delivered at the Certosa di Pontignano, Siena in May 1999, under the auspices of the Universita degli Studi di Siena I am grateful to participants at the meeting for discussion, and to later audiences at Bonn, Essen, Oxford and New York University I have retained the largely informal style of a conference talk Special thanks to Bill Brewer and Stephen Schiffer for valuable comments Some of the work on the present text was carried out while I held a Leverhulme Research Professorship, once again I thank the Leverhulme Trust for this invaluable support

[1] Bermúdez, *The Paradox of Self-Consciousness* (Cambridge MIT, 1998), especially chapter 3, Crane, "The Nonconceptual Content of Experience," in Crane, ed , *The Contents of Experience Essays on Perception* (New York Cambridge, 1992), pp 136-57, Dretske, *Naturalizing the Mind* (Cambridge MIT, 1995), especially chapter 1, pp 19ff , and earlier in *Knowledge and the Flow of Information* (Cambridge MIT, 1981), Evans, *The Varieties of Reference* (New York Oxford, 1982), see index entry for "Conceptual and Non-conceptual Content", Hurley, *Consciousness in Action* (Cambridge Harvard, 1998), especially chapter 4, my "Analogue Content," *Proceedings of the Aristotelian Society,* Supplementary Volume LX (1986) 1-17, and *A Study of Concepts* (Cambridge MIT, 1992), chapter 3, Tye, *Ten Problems of Consciousness* (Cambridge MIT, 1995), especially pp 137ff A rather different but related use of the notion of nonconceptual content is found in the work of Adrian Cussins see his paper "The Connectionist Construction of Concepts," in Margaret Boden, ed , *The Philosophy of Artificial Intelligence* (New York Oxford, 1990), pp 368-440

[2] McDowell, *Mind and World* (Cambridge Harvard, 1994), especially Lecture III and Afterword, Part II, Sedivy, "Must Conceptually Informed Perceptual Experience Involve Non-Conceptual Content?" *Canadian Journal of Philosophy,* XXVI (1996) 413-31, and Brewer, *Perception and Reason* (New York Oxford, 1999), especially chapter 5

0022-362X/01/9805/239-64

plausible as we proceed My specific aim here is to argue further for the thesis that perception does indeed have nonconceptual content[3]

I THE CLAIMS OF NONCONCEPTUAL CONTENT

The nonconceptual content of perception that is the topic of this paper is meant to be one species of representational content It is content that is evaluable as correct or as incorrect So we are not concerned here with whether there are nonrepresentational conscious properties of perceptions, what are sometimes called sensational properties The question I am now addressing concerns solely the class of representational contents of perception It is the issue of whether some of the representational content of perception is nonconceptual

The discussions of recent years have focused on the fine-grained representational content of experience When you look at the new Guggenheim Museum in Bilbao, or see a new abstract sculpture, or the face of a person, you see each of these objects as having a quite specific shape and size Similarly, you see them as having quite specific shades of colors, surface textures, and contours Equally, when you hear a musical tone, there is a sense in which you perceive its pitch You may not recognize the tone—you may not have absolute pitch—but you can discriminate that pitch from many, many others if you are asked to compare it with another This discrimination is based on the way the tone sounds to you All parties to these discussions have acknowledged the fine-grained character of this representational content What has been at issue is not its existence, but its character

We shall not do justice to the fine-grained phenomenology of experience if we restrict ourselves to those contents which can be built up by referring to the properties and relations that the perceived objects are represented by the experience as possessing We must, in describing the fine-grained phenomenology, make use of the notion of the *way* in which some property or relation is given in the experience The same shape can be perceived in two different ways, and the same holds for the shape properties, if we regard them as within the representational content of experience Ernst Mach's example of one and the same shape that can be perceived either as a square or as a regular diamond is a familiar example These are not different shapes The shape of an object need not alter when it moves, and an object can be perceived either as a square, or as a diamond, in either of the standard orientations relative to the perceiver

---

[3] I shall confine myself in this paper to the fundamental philosophical motivations for acknowledging nonconceptual content in perception, and the philosophical challenges such acknowledgment must address There are many further important issues about the nature and role of nonconceptual content which arise in the light of psychological experiments on perception, and on action discussion of them will have to be left to another occasion

The need to use the notion of the way in which something is perceived is by no means special to the perception of spatial properties and relations Suppose middle C and the F-sharp just above it are played simultaneously on a piano The interval thus sounded may be heard either as an augmented fourth, or as a diminished fifth In the former case, the higher tone of the interval is heard as the seventh of the presumed tonic scale [4] In the latter case, it is heard as the fourth of that scale In neither case need these music-theoretic descriptions be ones known to the perceiver, as part of his personal-level conceptual repertoire We use them in a plausible theory about the nature of the two ways in which the interval can be perceived It is the way itself, and not the materials of music theory, which enter the description of the subject's experience Even one and the same time interval may be perceived in two different ways It is perceived in one way, when it is perceived as the familiar time interval between your switching on your computer and its emitting the start-up sound It may be perceived in another way when the same interval is exactly that of the silence in a complete, but empty measure (bar) of a Haydn string quartet, in which case it may be heard as composed of two beats

So, in characterizing the fine-grained content of experience, we need the notion of the experience representing things or events or places or times, given in a certain way, as having certain properties or as standing in certain relations, also as given in a certain way Henceforth I use the phrase *the content of experience* to cover not only which objects, properties, and relations are perceived, but also the ways in which they are perceived The ways I have mentioned all contribute to the representational content of the experience That is, when something is perceived in one of these ways, the claim that the object really is the way it is experienced as being is one which has a correctness condition

Here, now, are five at least moderately intuitive claims about nonconceptual content which one might expect a believer in nonconceptual content to defend Some of these claims are, prima facie, incompatible with the position that the representational content of experience is exclusively conceptual Before we start discussing them, here is a brief, wholly dogmatic, statement of each of the claims

---

[4] On the musical case, see the fine contribution of Mark DeBellis, *Music and Conceptualization* (New York Cambridge, 1995), especially chapter 2

(1) The nonconceptual content of perceptual experience contributes to making available to a thinker various perceptually based concepts Only a thinker who has a perceptual experience with a certain kind of nonconceptual representational content can employ such perceptual-demonstrative concepts as *that shape, that texture, that interval of time* The nonconceptual content of experience also helps to make available such general, nondemonstrative recognitional concepts as the concept *regular-diamond shaped* and the concept *red* It also makes available more specific, but still general, recognitional concepts such as the concept *Cambridge blue* and the concept *yellow ochre* Perceptual states with nonconceptual content make these general concepts available to a thinker by providing the canonical, noninferential basis for the application of these concepts to things given in experience The nonconceptual contents that make available these various perceptually based conceptual contents cannot, however, be identified with any of the conceptual contents that they make available

These claims are all to be construed as constitutive What it is to have the perceptually based concepts is to be elucidated, philosophically, in terms of the relations of those concepts to the nonconceptual content of experience

(2) Experiences with finer-grained nonconceptual contents can also provide an empirical basis for the acquisition, and so enter the causal explanation of the learning, of such general concepts as *regular-diamond shaped*

(3) An experience with a certain nonconceptual content can make rational a judgment of a conceptual content suitably related to the nonconceptual content that the experience represents as correct

(4) Some of the nonconceptual content of our experience can be identical with the representational content of the experience of creatures that either possess no concepts, or possess only a set of concepts far more rudimentary than our own

(5) The nonconceptual content of experience can enter the explanation of features of intentional action There is an important discussion in psychology of whether this content, or its precursors, enter the explanation of action only via memory, rather than control concurrent action on-line [5] But all agree that it has an explanatory role in some cases

---

[5] For an illuminating overview, see A David Milner and Melvyn A Goodale, *The Visual Brain in Action* (New York Oxford, 1995)

The formulation of these five claims should make clear some of the wider ramifications of the otherwise apparently rather local issue of whether perception has a nonconceptual representational content.

We need to be very clear what we mean by 'conceptual'. I shall be taking it that conceptual content is content of a kind that can be the content of judgment and belief. Concepts are constituents of those intentional contents which can be the complete, truth-evaluable, contents of judgment and belief. Conceptual content and concepts I take to have identities conforming to, indeed answerable to, Gottlob Frege's criterion of identity for senses. Complete contents p and q are distinct if and only if it is possible for someone for whom the question arises rationally to judge that p without judging that q, and even while judging not-q. So the content 'This country is Italy' is distinct from the content 'This country is this country', the content 'The floor-plan is square' is distinct from 'The floor-plan is a regular diamond', the content 'Your lost pen is *there*' (pointing under a pile of papers) is distinct from 'Your lost pen is where you last used it'. Concepts C and D are distinct if and only if there is some completing content Σ such that the complete content $\Sigma(C)$ is distinct from the complete content $\Sigma(D)$, or, in other words, if and only if there is some content $\Sigma(C)$ such that someone for whom the question arises can rationally judge $\Sigma(C)$ without judging $\Sigma(D)$. So the pairs

this country	*Italy*
square	*regular diamond*
there (pointing appropriately)	*where you last used the pen*

are all pairs of distinct concepts by this criterion. As these examples show, concepts as intended here may be demonstrative or indexical.

On this approach, any further connections between concepts and, for instance, language have to be earned by further argument. If someone holds that a concept user must have a language in which he can express at least some of his concepts, that is a substantive, nondefinitional thesis that needs to be established. The same applies to the thesis that concepts, so characterized, conform to Evans's "generality constraint" (*op. cit.*).

It should, however, be uncontroversial that any content that can be expressed in language by the use of an indicative sentence, including sentences containing indexicals and demonstratives, will be a conceptual content. That follows in the presence of the less controversial premise that any utterance of an indicative sentence, in a given context, expresses a content that is of a kind that could also be the content of a belief or judgment.

The Fregean criterion as I have formulated it applies only to those contents which can be the content of judgment or belief. It would be a fair observation that the Fregean criterion as I have formulated it is an instance of a more general criterion, under which we can rightly say that there are distinct ways W_1 and W_2 in which a shape, or an interval, is given in perception simply because it is not obvious (even at the level of perception) that something given in way W_1 is the same shape, or interval, as something given in way W_2. So there certainly is some intuitive notion of "mode of presentation" that applies uncontroversially at the level of perceptual content. That more general notion of mode of presentation is not, however, the level at which the question here is posed. My question is whether the ways involved in perceptual content are the same as the conceptual contents that feature in the content of judgment and belief. The stipulations and terminology introduced so far do not settle this substantive question.

So much for minimal stage-setting. I turn now to argue in support of at least the claims (1) – (4), and, if it is agreed that they are true, to consider why they are true.[6]

II. DISTINCT FROM CONCEPTUAL CONTENT, BUT MAKING IT AVAILABLE

We can label as *conceptualists* those who claim that the representational content of experience is always conceptual. Conceptualists are committed to the falsity of the first claim (1), with its consequence that there is a level of representational content that cannot be elucidated in terms of conceptual content. Conceptualists have been tempted to say that the phenomena for whose description nonconceptual content has been invoked are better described by invoking conceptual contents containing perceptual-demonstrative concepts, or certain recognitional concepts. I take these two kinds of concepts in turn, and ask whether they can really meet the conceptualist's needs.

(a) Perceptual demonstratives. A perceptual-demonstrative concept *that shape* is a concept made available by an experience as of a particular shape. This perceptual-demonstrative concept will, in the context of its use in thought, refer to a shape, and the shape may be finely individuated. Similarly for the perceptual-demonstrative concepts *that shade, that sound, that taste*. Can the conceptualist say that it is these concepts which enter the representational content of expe-

---

[6] The discussion of (5) would involve more delving into the psychological literature than is feasible here. It is also significant that the case for the existence of nonconceptual content in perception can be made without reliance on any particular thesis relating it to action.

rience, and thereby fully capture the content that I have been claiming is nonconceptual?

An initial problem with this version of the perceptual-demonstrative route is the presence of the general concept in the perceptual demonstrative, as the concept *shape* features in *that shape* No one can enjoy perceptual states, or indeed any other conscious mental states, with the conceptual content *that shape* in their representational content unless he possesses the general concept *shape* It is, though, quite implausible that one must have that general concept in order to perceive objects as having various specific shapes Nor is it at all clear that two perceivers must have some more or less general predicative concept in order for them both to see some object as having the same specific shape One perceiver may think of a presented shape as *that rectangle*, the other perceiver may think of it as *that straight-sided figure* This difference in their thoughts need not prevent them from seeing it in exactly the same way

It is in any case not always required that a good, successfully referring perceptual demonstrative contain some general concept For those who think it must, a useful intermediate case that forces the softening of such a hard position is provided by the demonstrative *there* One difference between *there* and *that* is that the former must refer to a location, rather than anything present at that location But the demonstrative *there* does not do this by having the general concept of place as one of its constituents In fact, it is not at all obvious that a thinker, in order to have the rather unsophisticated capacity to think of a perceived location as *there*, must also have the general concept *place* or *location* in his conceptual repertoire Perhaps he must have the resources for introducing it, on the basis of concepts he already has but that is different from already deploying it

If this intermediate case is granted, it could hardly be denied that we could conceive of a family of demonstratives, modeled on *there* maybe *that-C* for the color apparently at a location, *that-T* for the surface texture there, and so forth But I do not think this would be a stable stopping point as a limit on the possibility of perceptual demonstratives unsupplemented with general concepts, nor as a suitable limit for the conceptualist position This is so, first, because we do not actually have such devices in our conceptual repertoire Our experience nevertheless still has the determinate, fine-grained character concerning textures, shades, and the rest all the same The fact that we could introduce such conceptual devices just serves to emphasize that what makes such devices available, namely, the rich representational content of experience, exists in advance of the conceptual apparatus it makes possible That makes it implausible that it can be identified with anything in that conceptual apparatus

Second, there seem to be cases in which a wholly unsupplemented perceptual demonstrative 'that' still secures reference in a suitable perceptual context You may think, when viewing a vase in a museum, 'That is beautiful', and be referring in thought to the color, not to the shape or the surface texture You may think, watching a Balinese dancer, 'That's beautiful', and be referring to the movement or the gesture rather than anything else Perhaps sophisticated thinkers can serve up general concepts if asked 'What kind of thing or event are you thinking of?' But it would be a stretch to insist that some general concept must have been entering the singular demonstrative component of their perception or their thought all along It seems, for instance, that someone could be introduced to the general concept *timbre*, applicable to sounds, by his first having an experience leading him to judge 'That's beautiful', referring specifically to the timbre of (say) a clarinet It may be that our listener only later applies the concept *timbre* to the instance he had already perceived and thought about ('That sound' could be too unspecific to capture what he experienced as beautiful)

The natural treatment of such cases is that the type of the perceptual demonstrative involved is individuated by a demonstrative element (which would expressed in an utterance by 'that')—an element to which I shall return—together with some particular way that enters the representational content of the perceptual experience that makes the whole perceptual demonstrative available The way in question is the way in which the shade, or movement, or shape, or whatever is referred to by the demonstrative, is perceived

For any given way in which a particular quality may be perceived, there is a specific kind such that that way is intrinsically a way for something of that kind to be perceived The way in which a shape is presented—for example, as a regular diamond—is intrinsically a way for a shape to be perceived, and not anything else The same applies to the way in which a texture, or a musical interval, is perceived It does not make any sense to suggest that a shape might be perceived in the way a musical interval is perceived In synaesthesia, the subject does not literally perceive a vowel sound in the way in which a color can be perceived On reflection, it should not be a surprise that for each way, there is a specific kind such that the way is intrinsically a way for something of that kind to be perceived Since the way contributes to a correctness condition, this general connection between ways and kinds is equivalent to this proposition that for an object to have the property a given way presents it as having, the object must meet a certain condition, and this condition is applicable only within objects of a given kind

This general connection between ways and kinds is important because it helps to contribute to an explanation of why a general concept is not always needed to supplement the 'that' in a perceptual demonstrative to secure manifest reference to one kind of quality of the perceived object or event, rather than another The kind of property referred to—a shape property rather than a sound property, say—is fixed by the perceptual way that contributes to the individuation of the perceptual demonstrative in question Provided that the demonstrative 'that' is linked to a given way, there will be a general kind such that the reference of the demonstrative must be of that general kind

Ways of being perceived are inherently general For instance, the shapes of two different objects may be perceived in exactly the same way, either on one occasion or on different occasions By contrast, perceptual-demonstrative reference to shapes themselves (to continue with that example) proceeds fundamentally via the shape of particular presented individuals When a thinker thinks 'That would make a good logo', referring to a shape, he is characteristically thinking of the shape as given in the perception of some particular object in his experience, and the truth conditions of his thought concern the perceived shape of that particular object Maybe one can refer perceptually demonstratively to the shape common to several objects, each of whose shape one perceives, and perceives to be the same for all of them—thus 'That shape, instantiated equally in all of those objects, would make a good logo' But this still involves a tie to particulars, several rather than just one This particularity, which seems distinctive of, and essential to, perceptual demonstratives, is one reason that one cannot simply identify the perceptual-demonstrative sense with the relevant way of being perceived The way, which is general, lacks the tie to particulars which is characteristic of the perceptual demonstrative

The link to a particular perceived individual that has the property demonstratively thought about is a characteristic of the demonstrative element in the perceptual-demonstrative concept It would not be right to regard a perceptual-demonstrative concept *that shape* as thought on a particular occasion as built up from a way in which a shape may be perceived plus some entirely generic constituent The link to a particular, given in perception or experience, must be captured if we are to characterize fully the perceptual-demonstrative concept used on this occasion

It is important to distinguish between ways of being perceived and ways of having a certain generic quality A way in which a shape property may be perceived is to be sharply distinguished from a way of being shaped A way of being shaped has to do with shapes themselves—it is a way of occupying space—and does not have to do with the way in which shapes are perceived On the natural, unforced way of understanding the notion of a way of being shaped, the property of being a square and the property of being a regular diamond are actually the *same* way of being shaped This means that one could not individuate perceptual concepts, whether fine grained or rough grained, in terms of which way of being shaped they are concepts of If one uses 'way' in this territory for 'way of being shaped', or 'way of being colored', or more generally 'way of having some generic kind of perceptible property or relation', one will need additionally something else to capture the different ways in which these ways of being shaped may be perceived

Although a way contributes to a correctness condition, it is important that which object is presented in a given way is not simply a matter of the object's fitting that way A distant aircraft in the sky may be presented as being in a certain direction It may not in fact be in that direction, because the light rays are passing through refracting bodies of differentially heated air When an object is presented in a given way in perception, which object is presented in that way is at least partly a matter of causation, as H P Grice[7] argued It is a consequence of this point that one cannot regard ways as individuated by the condition for them to refer to some object or property It is, however, a highly plausible thesis about each Fregean sense (including perceptual demonstratives) that it can be individuated by the condition for it to refer to a given object or property [8] If this plausible thesis is correct, then we have a further argument that ways are not Fregean senses

This distinctness of ways from senses is further confirmed by the natural explication of the conditions for a perceptual-demonstrative concept to refer to a given object Those conditions concern the causal relations of an object to a way in which the object is given in perception Such conditions take for granted the notion of a way, and then draw upon it in elucidating the condition for the demonstrative to refer to a given object

---

[7] "The Causal Theory of Perception," *Proceedings of the Aristotelian Society*, Supplementary Volume xxxv (1961) 121-52

[8] See Michael Dummett, *The Interpretation of Frege's Philosophy* (Cambridge Harvard, 1981), chapter 3, pp 45ff

These remarks against treating fine-grained perceptual content as explicable in terms of perceptual-demonstrative concepts should not be taken to disparage the very idea of a fine-grained perceptual demonstrative. There are perceptual-demonstrative ways of thinking of presented shapes, colors, sounds, and the rest. Perception of an object's shape may make available a way of thinking of that particular shape, as 'that shape'. This is a way of thinking that also makes it possible to wonder or to speculate whether some other object has that very shape, the shape one perceives the presented object to have.

It is constitutive of perceptual demonstratives that they can be used by the thinker only while the perceptual experience that makes them available continues. McDowell suggests that such a dependence on the occurrence of the experience "would cast doubt on its being recognizable as a conceptual capacity at all" (*op. cit.*, p. 57). This may be one of the reasons he moves to endorse recognitional concepts as the correct treatment of perceptual content. Of experience-dependent demonstratives, he says that they look "like Wittgenstein's case of the person who says 'I know how tall I am', putting his hand on top of his head to prove it" (*op. cit.*, p. 57).[9] But the person in Ludwig Wittgenstein's example does not, by saying or thinking '*this tall*' come to know his height in meters or any other unit, which is what knowing how tall one is requires on the usual understanding. It is not even clear that such a person thereby (I emphasize thereby) comes to know how tall he is, in the sense that he could indicate which of the markings on a wall on the opposite side of the room is roughly his height. By contrast, a perceptual-demonstrative thought latches on to a magnitude, or shade, or color, only if that magnitude, or shade, or color, is itself given in the experience that makes the perceptual-demonstrative concept available. There is no possibility of making perceptual-demonstrative reference to a magnitude, or shade, or color, and not knowing what magnitude, or shade, or color it is one is thinking about. Such perceptual-demonstrative reference is the most fundamental way of knowing what magnitude, or shade, or color is in question.

We should also distinguish here the conditions for making demonstrative reference in language from the conditions for employing a perceptual-demonstrative concept. The person who utters 'that table', pointing behind himself to a table he is not seeing, certainly refers to a table, but he has no perceptual-demonstrative concept of it A similar point could be made for an utterance of 'that shade', pointing to the wall behind oneself, without looking at it There are also important background conditions that must be met for perceptual-demonstrative reference in thought to succeed Stroboscopic lighting, undetected movement, and much else can undermine the holding of these background conditions Perceptual-demonstrative reference can fail even when it seems to the thinker to succeed (These background conditions will be all the more important to one, like myself, who thinks that perceptual demonstratives and observational concepts can be individuated in terms of the conditions under which contents containing them can be not merely rationally judged, but *known* [10]) But there are plenty of cases in which the required background conditions are fulfilled, and there can be perceptual-demonstrative reference to finely-individuated properties and relations My point has not at all been that such perceptual demonstratives are nonexistent Rather, my point is that they are themselves individuated in part in terms of ways in which properties, magnitudes, and relations are given in experience, and so cannot be used to elucidate the nature of such ways of being experienced

---

[9] His reference is to Wittgenstein's *Philosophical Investigations*, G. E. M. Anscombe, trans. (Oxford: Blackwell, 1958, second edition), § 279.

(b) Recognitional concepts The other position tempting to some conceptualists is that the fine-grained representational content of experience is to be captured not by perceptual-demonstrative concepts, but by fine-grained recognitional concepts This is McDowell's view in *Mind and World*, and it squares with his most explicit recent discussion of the issues [11] McDowell writes in *Mind and World* "Why not say that one is equipped to embrace shades of colour within one's conceptual thinking with the very same determinateness with which they are presented in one's visual experience, so that one's concepts can capture colours no less sharply than one's experience presents them?" (*op cit*, p 56) Soon after in the same book he makes clear "What is in play here is a recognitional capacity " (*op cit*, p 57) The recognitional capacity associated with the recognitional concept persists, if only briefly, after an experience which is of the

---

[10] See my *Being Known* (New York Oxford, 1999), chapter 2

[11] "Replies to Commentators," *Philosophy and Phenomenological Research*, LVIII, 2 (June 1998) 403-31 In the part of these "Replies" which deals with me, McDowell writes always of demonstratives, rather than recognitional concepts But it is important that for McDowell in his Reply, "'that way' in 'It looks that way' is predicative, not substantival Expressions at this place in this construction express, rather than refer to, ways things can look" (p 418) On McDowell's theory, expressions in that place express recognitional concepts This is fully in accord with his position in *Mind and World*—see p 57

kind recognized (*op cit*, p 57) McDowell seeks to capture part of the fine-grained spatial content of experience with a conceptual content of the following sort, which he would say is represented as correct by the experience

This is shaped *R*

Here *R* expresses, for McDowell, a recognitional concept of a way of being shaped Similarly for color, on McDowell's view we have the conceptual content

This is colored *S*

in the representational content of the experience, where *S* is a recognitional concept of a fine-sliced shade

If it was wrong to have the general concept *shape* in the demonstrative attempts to capture the fine-grained content of experience, it seems equally wrong to have the general relational concept *shaped* ______ in the representational content under this conceptualist treatment The same applies to the concept *colored* ______ I shall not, however, trade on this point, and shall take it that McDowell could use precisely the points I made earlier about the absence of a need for supplementation by a general concept He could move simply to propose the conceptual contents 'This is *R*', 'This is *S*' as the fine-grained representational content of experience The recognitional capacities underlying these recognitional concepts persist, "possibly for quite a short period," according to McDowell in *Mind and World* (*op cit*, p 172), where he holds that they can be employed in the absence of perception of the finely-sliced properties to which they refer

As Diana Raffman[12] observes, it is well known that perceptually-based discrimination of properties far outreaches memory and identification of those properties Recognitional capacities are by their very nature constrained by memory capacities I would add to Raffman's point only two observations First, even if memory were as finely discriminating as perception, that still would not make it right to regard recognitional concepts as entering the representational content of experience There cannot be recognition when the perceptual property is encountered for the first time in a given way A first-time experience of a property nevertheless still has a specific, fine-grained representational content And second, in those counterfactual circumstances in which memory were as discriminating as perception, the experience would still have a specific content which could exist even if the recognitional capacity did not The memory structures required for the recognitional concept and its corresponding recognitional capacity, and their subpersonal underpinnings, could all be absent, and yet the experience still have a fine-grained representational content The specific fine-grained content would show up in matching tests that do not need to rely on memory, as opposed to current perception This modal point applies however short the period for which the recognitional capacity is said to exist

---

[12] "On the Persistence of Phenomenology," in Thomas Metzinger, ed , *Conscious Experience* (Paderborn Schoeningh, 1995), pp 293-308, pp 294ff

It seems to me, then, that neither perceptual demonstratives nor recognitional concepts can capture the representational content of perception But however convincing it may be, candidate concept by candidate concept, that each candidate concept cannot capture perceptual content, it is reasonable to ask here for more theoretical explanation Is there some deeper, and more general, reason why perceptual content cannot be captured by any kind of conceptual content? I shall return to this, when we are in a better position to characterize more fully the nature of the distinction between conceptual and nonconceptual content

III LEARNING OBSERVATIONAL CONCEPTS

Now let us take the phenomenon of learning a new observational concept from instances of it that are perceived when one is taught, or otherwise acquires, the concept Consider the representational content of the subject's experience when she is learning the concept *pyramid,* considered as an observational shape concept

On the one hand, the representational content of this experience, in the case of learning from a positive instance, must be sufficient for someone rationally to apply the concept—must entitle her to apply the concept—when experience is being taken at face value If that were not so, it would not be an observational concept after all Yet on the other hand, this representational content cannot include the concept *pyramid,* for that would require the learner already to possess the concept The learner could have such experiences only if the lesson were unnecessary

The natural solution to this last quandary is to acknowledge that there is such a thing as having an experience of something as being pyramid shaped that does not involve already having the concept of being pyramid shaped What such an experience will have is a nonconceptual content which, if correct, is sufficient for something's falling under the observational concept *pyramid*

In fact, that is to understate the case Not merely is the holding of the nonconceptual content represented as correct in the positive cases involved in learning sufficient for the object in question to be a pyramid. The experiences from which, as presenting positive instances, one can rationally learn the concept, are those which have nonconceptual representational contents of the sort mentioned in the very condition for possessing the observational concept that is being learned. There is a specific kind of nonconceptual content of experience to which one's judgments 'That's a regular diamond' must be rationally sensitive if one is to possess the observational concept *regular diamond*. There is a different specific kind of nonconceptual content of experience to which one's judgment 'That's a square' must be rationally sensitive if one is to possess the observational concept *square*. There is a specific kind of nonconceptual content of experience to which one must be sensitive if one is to acquire a recognitional concept for the musical interval of an augmented fourth. It is a different kind from that to which sensitivity is required in possessing a recognitional concept for a diminished fifth, and so forth.[13]

This rational sensitivity is not a matter of inference, even though the occurrence of an experience with a certain nonconceptual content does make rational a perceptual judgment with a conceptualized content. It is intuitively wrong to classify the case as one of inference. It also could not be correct, if the perceptual content is nonconceptual: inferential relations can hold only between (states with) conceptual contents. We can, as reflective thinkers, indeed introduce ways of thinking about the nonconceptual ways in which things are given. But this is built on a more primitive level of rational sensitivity to the experience's nonconceptual content itself. Such reflective thought is not necessary for having the capacity to make perceptual judgments about one's environment.

IV. MAKING RATIONAL TRANSITIONS TO PERCEPTUAL JUDGMENTS

I have emphasized that the way in which some thing, or property, or relation is given in the nonconceptual content of an experience is something that contributes to what it is like to have that experience. These ways which feature in nonconceptual content are then at the conscious, personal level, and are not merely subpersonal. As features of the subjective experience, their presence can entitle a thinker to make a particular judgment, or to form a certain belief.

---

[13] For some observational concepts, the kind may be amodal, and so instantiated in experiences in more than one modality. This is plausible for *straight* and *right-angled*, construed as observational concepts.

A thinker can be rational in making a transition from an experience with a certain nonconceptual representational content to a judgment with a certain conceptual content, in particular in making a transition to judging a content in which an observational concept is predicated of presented objects or events Such a transition is rational when the thinker is entitled to take her experience at face value, and when the observational concept is individuated in part at least as one that the thinker must be willing to judge when experience has a certain kind of nonconceptual representational content (and is being taken at face value) Consider Mach's example of a closed right-angled figure formed from four straight lines, and whose symmetry about the bisectors of its opposite angles is perceived Such an experience can make rational the judgment *That's a regular diamond* It does not make rational the judgment *That's a square* That latter judgment is made rational rather by a different nonconceptual content, in which the perceived symmetry of the closed figure is a symmetry about the bisectors of its sides Which symmetries are perceived in each case does not need to be something known to, or even conceptualized by, the perceiver

The nonconceptual content in question throughout this paper is nonconceptual representational content, with a correctness condition Consider any case in which an observational concept *F* is plausibly individuated partly by its relations to a particular sort of nonconceptual content It is then a philosophical task to show that the holding of the correctness condition for the nonconceptual content in question ensures the holding of the correctness condition for the conceptual content *That's F* In our simple spatial cases, one could show this by showing that the shape which something must be when it is presented (veridically) in an experience with a certain nonconceptual content is a shape which is sufficient for the observational concept *is square* to apply to it

This is an instance of a task that exists for any rational transition that involves a concept or concepts at some point or other The task exists equally for the case of logical transitions If a premise gives a good reason for accepting a certain conclusion, because of the logical form of the inference, there must be some explanation, in the semantics of the logical constants involved in the transition, of why this is a valid transition That is why we need a semantics, and not merely a proof theory, for a logic (Even in the special case in which the semantics is itself proof theoretic, the need for a semantical theory is still being granted) The ordinary thinker does not need to know what the explanation of the validity of a particular transition is, in either the logical or the perceptual case But the explanation must exist if the transition is to be valid Only so is there an explanation of why it is truth preserving, and thereby of how it can lead to new knowledge

My own view is that it is always a task of philosophy to explain why a transition is a good one This requirement, in my view, applies to the conceptual-to-conceptual case of logic, to the nonconceptual (but still representational)-to-conceptual case of perceptual judgment, and to the arguably wholly nonrepresentational-to-conceptual transition from the occurrence of a pain in a subject to his conceptualized self-ascription of pain [14]

The general position on nonconceptual content that I have outlined so far may seem intuitive, but it has been subjected to a philosophical critique by McDowell His most fundamental objection is that experiences can have a rational bearing on thought only if their content is conceptual "The point of the claim that experience involves conceptual capacities is that it enables us to credit experiences with a rational bearing on empirical thinking" (*op cit*, p 52) This claim needs an argument, and McDowell's argument for it in *Mind and World* is that rational thinking involves the possibility of scrutinizing the relations between experience and judgment Nothing is outside the limits of rational self-scrutiny (*op cit*, p 52), even if, after scrutinizing the case in question, we decide to endorse it McDowell's view is that it is hard to see how this scrutiny would be possible unless experience has only conceptual content (*op cit*, pp 52-53)

That last step seems to me a nonsequitur I accept that we can rationally scrutinize any instance of any putative reason-giving relation, including the relation between experience and judgment McDowell's concern is that on a theory like mine, the subject cannot gain the required "comprehensive view of the two contents"—that of the experience and that of the judgment (*op cit*, p 166) On my treatment, however, a thinker can ask 'Is something's looking that way a reason for judging that it's square?', for instance On the approach I advocate, 'that way', in this particular occurrence, refers demonstratively to a way in which something can be perceived The reference itself is made *by* something conceptual demonstrative concepts can enter conceptual contents There is no requirement that the reference *of* the demonstrative be conceptualized (I shall return to such uses of 'that way' in a few paragraphs) So thought can scrutinize and evaluate the relations between nonconceptual and conceptual contents, and obtain a comprehensive view of both On McDowell's theory, of course, the demonstrative 'that way' expresses a sense (a recognitional concept), which does on his view enter the representational content of experience I have rejected that treatment on other grounds—the required recognitional capacities do not exist, and do not need to exist for the experience to have a fine-grained content But quite apart from that, the mere possibility of the treatment I have offered shows that it is a nonsequitur to move from an agreed premise of the required scrutinizability in conceptual thought of the rational relations between experience and judgment to the conclusion that experience itself must have only conceptual representational content As long as we can think about the nonconceptual representational content, as we certainly can, we can have the required scrutinizability

---

[14] I hope this and the preceding paragraphs clarify my intentions at around page 80 of *A Study of Concepts*, about which McDowell expresses some puzzlement "it is hard to see much in the way of a further issue about how the reason [supplied by an experience with an allegedly nonconceptual content] can be a *good* one" ("Replies," p 418)

McDowell also holds that there is a requirement of articulability of reasons, including perceptual reasons, and that this too supports his view that the representational content of experience must be wholly conceptual "[Peacocke] has to sever the tie between reasons for which a subject thinks as she does and reason she can give for thinking that way Reasons that the subject can give, in so far as they are articulable, must be within the space of concepts" (*op cit*, p 165) I am not at all sure myself that there is such a requirement of articulability But if there is, it seems to me that the use of the demonstrative 'that way', construed in the way it is on my approach, would allow the friend of nonconceptual content to meet any such demand 'I believe it's square because it looks that way' is an articulation of the subject's reasons Once again, the conceptual character of the conceptual constituent 'that way' must be sharply distinguished from the nonconceptual character of its reference, a nonconceptual way in which something is perceived

I would invoke demonstrative ways of thinking of ways in addressing a closely related objection of Brewer's (*op cit*, pp 166-69) He argues that a thinker must be capable of recognizing her reasons as such, that is, as reasons This demand does not, Brewer argues, intrinsically raise difficulties of principle in the case of a mental state that gives reasons, provided that the state has conceptual content But in the case of a mental state with an allegedly nonconceptual content, "mere possession of the state involved in having such a 'reason' is explicitly insufficient for the subject's understanding of the proposition whose association with this state grounds its putative status as his reason for doing what he does The only alternative seems to be that he must have some second-order knowledge of the relation between mental states of the type in question and the truth of the belief for which he thereby, and only instrumentally, recognizes his having a reason" (*op cit*, pp 167-68) Brewer says that the nonconceptualist is "condemned to follow the discredited classical foundationalists and coherentists" (*op cit*, p 166) In short, his position is that the nonconceptualist can provide for the recognition of his own reasons only by arguing as follows

(i) This state is *F*
(ii) Anything that is *F* is a reason for believing that *p*
(iii) Hence I have reason to believe that *p* (*op cit*, p 168)

Brewer adds that "what stands in place of '*F*' will have, at best, to be hideously complex" (*op cit*, p 168)

I agree with Brewer that, if the nonconceptualist could account for the recognition of the reason that his experience gives him for making a perceptual judgment only in this second-order, instrumental fashion, the position would be unacceptable But I maintain that the nonconceptualist's position does not have the consequence Brewer claims Here is a way that a thinker can come to recognize a state with nonconceptual representational content as giving reasons, without taking the second-order, instrumentalist route The perceptual experience represents some presented object or event as having some property, given in a certain way *W* So the experience has a content (if we put it, conveniently but inessentially for the present point, in object-dependent form)

(iv) *x*, given in way *s*, has property *P*, given in way *W*

The way *s* in which the object is given will capture its apparent distance and direction, egocentrically characterized, from the perceiver[15] Now suppose this way *W* is, for example, one of the nonconceptual ways in which a shape can be given, and is mentioned as sufficient for being square in the possession condition for the concept *square* On my account, the subject is then entitled to move from (iv) to the conceptualized content

(v) That object [given in way *s*] is square

---

[15] No doubt the way *W* could also be said in some pre-theoretic sense to be part of the way the object is given in perception But it is important to distinguish between those ways which help to determine *which* object is perceived from those which do not *s* is in the former class—though as emphasized earlier, this determination is not a matter of fitting

Similarly for any other way whose presence in an experience gives good reason for applying an observational concept

Now, what would be involved in a subject's reflecting on these transitions, and in coming to appreciate his reasons as reasons? (iv) and (v) do not seem to be hideously complex, complexity is not the heart of the issue But there are two interesting capacities a thinker must have if she is to be capable of appreciating both the occurrence, and the rationality, of her transition from (iv) to (v)

First, our reflective subject must have a special way of thinking of the way *W* There seems to be a phenomenon here that is a partial analogue, for the relations between nonconceptual contents and demonstrative concepts of them, of the relations that hold between concepts and canonical concepts of those concepts—"conceptual redeployment," as I called it [16] When some property is given to a thinker in perception in way *W*, then, if that thinker is sufficiently reflective, there seems to be a conceptual way C_w of thinking of that way *W*, where this conceptual way C_w seems to be made available to the thinker precisely by her enjoying an experience whose nonconceptual content includes that way *W* Presumably, this relation of 'being made available by' consists at least in part in a certain canonical sensitivity of judgments involving the concept C_w to evidence about the way *W* which the subject has precisely because she is experiencing some property as given in way *W*

Second, in appreciating that the occurrence of an experience with the representational content (iv) entitles her to judge the content (v), our thinker will be displaying a sensitivity to a particular boundary This is the boundary between those states which are, and those states which are not, mentioned in the possession condition for a given concept such as *square* Our thinker will have moved from the ground level of satisfying a possession condition to the level of thinking about it One, perhaps the basic, way to make this transition is to ask questions that are in the first instance not about concepts, but about the world Such a question might be 'Would something's looking that way *W* give reason to think it is square?' One can answer such questions by drawing on one's ordinary, ground-level abilities to react rationally to one's perceptual states in coming to make judgments about the world Someone who satisfies a possession condition can clamber her way up to reach some understanding of what that possession condition is

---

[16] See my *Being Known*, sections 5 5-5 6

This description of the transition from perception to judgment, and of the subject's appreciation thereof, is not second order Nor is it of the highly problematic kinds in the sorts of classical foundationalist and coherentist writings Brewer rightly criticizes (iv) and (v) are not themselves about mental states at all It is true that (iv) becomes pertinent because of its relation to the content of a mental state But that will equally hold on a conceptualist's account of the reason relation and of appreciation of it as a relation of reason giving The content is conceptual, according to that theorist, and stands in various reason-giving relations appreciable by the thinker But that content is relevant only because it is the representational content, on the conceptualist's view, of the perceptual experience in question That it is relevant only because of that relation does not make the conceptualist's view second order, or instrumentalist Equally the corresponding feature of (iv) and (v) cannot make the nonconceptualist's account second order or instrumentalist either It is the relation between (iv) and (v) which is crucial in the account of the reason-giving relation, and appreciation of it as such, and that does not involve anything second order

A full theoretical, philosophical account of why the transition from (iv) to (v) is valid would of course involve some philosophical theory The theory would be part of a general theory of content A full theoretical account would equally be required in the case of logical transitions between purely conceptual contents The philosophical theory does not need to be known to the ordinary thinker in either of these rather different cases In both cases, there is a willingness to make transitions between certain contents, which is constitutive of possession of the observational concept or the logical concept, respectively On my account, in the former case the content of the entitling or reason-giving state is nonconceptual, and in the latter it is conceptual The ordinary thinker can appreciate his reasons as reasons in either case In either case, he may or may not go further in reflecting on the nature of the good reasons involved

A different question about this approach does not query the distinctions I have drawn, but raises at least a problem about how the issues have been framed [17] Surely, I would not object to a regimentation of the content of a conceptual state in which, for instance, a recognitional concept for a particular size of diamond shape were captured by the very nonconceptual content mentioned in its possession condition? But if we can use nonconceptual content in capturing

[17] Here I am indebted to Stephen Schiffer

conceptual content, has the issue with the conceptualist been properly joined when we take it as the question of whether experience has only conceptual content? I agree that we could employ such a means of capturing a recognitional concept But if we do use the apparatus of nonconceptual content in characterizing conceptual states, we must distinguish sharply between the kinds of relations in which a state must stand in order to possess the original nonconceptual contents, and those in which it must stand in order to possess those conceptual contents which are, under this proposal, captured by nonconceptual contents The arguments of the next section will further distinguish these two sets of relations

V KINDS OF CONTENT, ANIMAL PERCEPTION, AND OBJECTIVITY

Cats, dogs, and animals of many other species, as well as human infants, perceive the world, even though their conceptual repertoire is limited, and perhaps even nonexistent These perceptions are subserved by perceptual organs, and in the case of higher species, subserved by brain structures similar in significant respects to those which subserve mature human perception By the "hard line" on animal perception, I mean the thesis that none of the conscious perceptual states with representational content enjoyed by mature humans can be enjoyed by nonlinguistic animals without concepts, or with only minimal conceptual capacities By the "soft line," I mean simply the denial of the hard line So the soft line says that some of the conscious perceptual states with representational content enjoyed by mature humans can be enjoyed by nonlinguistic animals without concepts, or with only minimal conceptual capacities

For what it is worth, pretheoretical intuition seems to find the hard line too hard to accept The hard line entails that the following cannot be literally true that the animal has a visual experience as of a surface at a certain orientation, and at a certain distance and direction from itself, in exactly the same sense in which an adult human can have a visual experience with that as part of its content Abandoning all pretense at unbiased terminology, I shall call the conclusion that that cannot be literally true *the unintuitive conclusion* The soft line, which prima facie at least is not committed to the unintuitive conclusion, is naturally developed hand-in-hand with a theory of nonconceptual representational content According to the soft line, it is nonconceptual representational content which can be common to visual experiences which both you and a mere animal may enjoy

Those who take the hard line need not deny that animals have some perceptual sensitivity to their environment, a sensitivity which explains their actions. McDowell, who takes the hard line, insists that animals have such a perceptual sensitivity. "It is a plain fact that we share perception with mere animals" (*op. cit.*, p. 114).[18] Nor does McDowell have in mind, for instance, only some kind of informational sensitivity that would make animals automatons of the sort which René Descartes apparently believed them to be. In *Mind and World,* McDowell speaks of "proto-subjectivity" (*op. cit.*, p. 117). He says that there is a legitimate kind of talk of what the features of the environment are for an animal, which "expresses an analogue to the notion of subjectivity, close enough to ensure that there is no Cartesian automatism in our picture" (*op. cit.*, p. 116). All the same, it must be only an analogue on McDowell's view. It is, on his account, not literally true that the mere animal has a visual experience as of a surface at a certain distance and direction in exactly the same sense that mature concept-using humans do.

It is certainly a necessary condition of being reasonable in rejecting the unintuitive conclusion that one address the arguments that in perceptual experience, there is only conceptual, and not nonconceptual, content. But if we can develop a theory of nonconceptual representational content, and answer the conceptualist's objections to it, as I have been trying to do, then much of the pressure to adopt the unintuitive conclusion is substantially relieved.

Although I think, however, that the unintuitive conclusion is indeed to be rejected and the soft line is right, the truth in this area seems to me much more interesting than those somewhat flat conclusions suggest. It seems to me that there is a good, Kantian point in McDowell which should not get lost in the endorsement of the soft line. There is something very plausible in McDowell's Kantian position that "the objective world is present only to a self-conscious subject, a subject who can ascribe experiences to herself. It is the power of conceptual thinking that brings both the world and the self into view. Creatures without conceptual capacities lack self-consciousness and—this is part of the same package—experience of objective reality" (*op. cit.*, p. 114). Although of course a creature with perceptions with nonconceptual representational contents has states whose correctness conditions concern the objective world, it is a question whether a creature without the specific conceptual apparatus McDowell mentions would be conceiving—would even have the resources for conceiving—this world *as* objective So the question arises Is there some way we can consistently acknowledge this point while also maintaining the soft line? I shall be arguing that there is—but to do so we need some deeper theory of the nature of the distinction between conceptual and nonconceptual content

---

[18] See also pp. 50, 64-65, and Lecture VI, section 4-7.

I first draw a distinction within the possible grounds, or reason-giving states, which lead, in a given subject's psychological economy, to acceptance of a given kind of content We can distinguish those grounds being *exhaustive*, for a given content, and their being *canonical* Consider the minimal possible ways a creature capable of perceptual states with nonconceptual content may come to accept of three perceived objects *a*, *b*, and *c* that they are in a straight line, or may come to reject that same content The content may be accepted on the basis of a perception with a nonconceptual representational content to that effect Such a creature may also store this information, with or without memory images of the original perception The content may come to be rejected if, in moving or being moved around, the subject does not have perceptions of the sort which, from other angles, should be produced by *a*, *b*, and *c* if they are in a straight line Now suppose such perceptual and cognitive-map based ways of coming to accept or reject exhaust the subject's ways of coming to accept or reject such contents (or at least, exhaust those which are distinctive of those contents, if this subject is capable of engaging in more general modes of reasoning) Such exhaustiveness of the perceptual-cum-cognitive-map based ways of coming to accept or reject those contents contrasts sharply with the case of possession of the concept of being in a straight line For a thinker with the concept *collinear*, any grounds for thinking that three things stand in the same relation as is perceptibly instantiated when the thinker sees three things to be collinear are equally grounds for judging those first three things to be collinear There are no restrictions on what those kinds of grounds may be What those grounds may be is something to be discovered The grounds are in no way restricted to what can be attained through perception and the use of a cognitive map These additional grounds need not be anything particularly sophisticated Even the child who moves some straight-edged physical object to check whether three objects are in a straight line, when he cannot tell this, for one reason or another, just by looking, is manifesting a sensitivity to this open-endedness of grounds

The presence in the thinker's thought of possible additional grounds is entirely consistent with perception having some canonical status in the individuation of the concept *collinear*, considered as an observational concept Some grounds can have a privileged status, consistently with there being no restrictions in principle in a thinker's thought on the range of possible grounds The fact that certain grounds are canonical does not mean that they are exhaustive We must distinguish the exhaustive from the (merely) canonical

It is important to notice two points about the distinction between the exhaustive and the canonical First, when a certain body of grounds is not exhaustive in principle in a thinker's thought, the thinker is tacitly operating with the notion of the same property, or same relation, in the world It is the same relation that can be detected, on the one hand, by perception with the use of cognitive maps and, on the other, by the additional means—the ruler, or whatever else the ingenious thinker may come up with The thinker for whom grounds involving perception-cum-cognitive maps are not exhaustive for acceptance or rejection of a given kind of content must be capable of kinds of reasoning that involve grasp of the notion of sameness of property and sameness of relation The same holds for the case of sameness of objects, if we apply these points to the distinction between genuine concepts of objects, and the ways objects are given in the nonconceptual content of perception

The second point to notice about this distinction is that it has to do with the nature of objectivity It is a consequence of the idea of things being objectively thus-and-so, of its being objectively the case that these objects stand in this relation, that their doing so is something that can be known about in many different ways It is not restricted in principle to any one mode of access We do not have fully objective thought unless a thinker manifests some grasp of a conception of objectivity with this character

It is here that my discussion connects up somewhat more closely with McDowell, despite our major differences For he insists that only with conceptual contents does one have a certain kind of objectivity—and I have recently been asserting just that The reasons are not, of course, precisely the same I have not, for instance, committed myself to the view that there is properly objective thought, thought about something conceived as objective, only if the subject actually has the concept of perceptual experience All the same, I am in agreement with McDowell that the minimal requirements for thought about an objective world that is in some way conceived as objective are not met simply by the possession of perceptual states with nonconceptual representational content

A similar point holds also for a subject's conception of himself, herself (or perhaps, in the case of some mere animals, itself) The reasoning is essentially the same it relies again on the distinction between the exhaustive and the canonical Suppose a thinker's only way of establishing contents about himself is by the use of the following

(a) By perception—'I am in front of a lake', 'I am under a tree'
(b) By memory— 'I was in front of a lake', 'I was under a tree'
(c) By integrating the deliverances of perception into a cognitive map and extracting the consequences thereof—'There's a wood behind me', 'There's a lake an afternoon's walk from the wood in that direction'

That is, for this thinker perception, memory, and the operation of his cognitive map in the ways illustrated above exhaust his ways of establishing things about himself[19] This is not yet to have a full conception of himself, as an object in the objective order of things In particular, it involves no conception of what it would be for something to be true of himself that is not establishable as true by the above methods Contents not so establishable include 'When I was asleep, I rolled over four times', 'There are times I existed of which I know nothing' Full objective thought about oneself must, as in the case of conceptual thought about properties and relations, involve some conception of oneself as an object truths about which are not exhausted by those which can be established in a certain restricted set of ways The subject who is oneself may exist in times, places, and circumstances to which one's current means of establishing first-person contents cannot reach

I conjecture that this distinction, in point of grasp of objectivity, between the minimal requirements for having states with nonconceptual contents, and what is involved in conceptual content, is a deeper reason why perceptual content cannot be explained in terms of conceptual content The most primitive aspects of representational content in perception, which our subjective experience shares with the mere animals, do not involve the grasp of objectivity required for conceptual content This is one of the reasons that trying to treat all perceptual content as conceptual involves an overascription We should always distinguish between content that is objective, and content which is not only objective, but which is also conceived of as objective

CHRISTOPHER PEACOCKE

New York University

---

[19] On nonconceptual self-consciousness, see Hurley, chapter 4

BOOK REVIEWS

New Horizons in the Study of Language and Mind. NOAM CHOMSKY, with a Forward by NEIL SMITH. New York: Cambridge University Press, 2000. xvi + 230 p. Cloth $54.95, paper $19.95.

Here are seven essays that describe and deplore a philosophical double standard that respects the methods and results of physics, chemistry, and biology but not the methods and results of linguistics and other sciences of the mind. One sign of the double standard is that, while hardly anyone thinks one can do philosophy of physics without knowing physics, it is all too common to think one can do philosophy of language without knowing linguistics.

Noam Chomsky is, of course, the leading figure in contemporary linguistics. Starting in the 1950s, his development of generative grammar was an important factor in the shift from behavioristic to cognitive approaches to language and mind. Chomsky's approach takes the goal of linguistics to be to characterize the human faculty of language, noting its differences from the human faculties for general problem solving and science. As Chomsky and other linguists tried to give explicit characterizations of the competence of a speaker of a language like English, it became clear that a child learning a language simply does not have the sort of evidence available that would enable it to learn the relevant principles from scratch. There is a "poverty of the stimulus." The child must be prepared to acquire a language with these principles in a way that it is not prepared to acquire the principles of, say, physics or quantification theory.

It is clear that normal children acquire a language that reflects their particular linguistic environment. A child brought up in Japan acquires a version of Japanese. The same child brought up in Brazil acquires a version of Portuguese. So, these languages must in some sense reflect some of the same underlying innate principles.

Further reflection along these lines and a great deal of empirical study of particular languages has led to the "principles and parameters" framework which has dominated linguistics in the last few decades. The idea is that languages are basically the same in structure, up to certain parameters—for example, whether the head of a phrase goes at the beginning of a phrase or at the end. Children do not have to learn the basic principles, they only need to set the parameters. Linguistics aims at stating the basic principles and parameters by considering how languages differ in certain more or less

0022-362X/01/9805/265-69

subtle respects. The result of this approach has been a truly amazing outpouring of discoveries about how languages are the same yet different.

More recently, there have been attempts to try to explain some of the basic principles on the assumption that the language faculty is close to an ideal engineering solution to a problem of connecting the language faculty with the cognitive system and the articulatory perceptual system. This "minimalist program" remains highly speculative, but whether of not it succeeds, contemporary linguistics as a whole has been a tremendous success story, the most successful of the cognitive sciences.

I. SCIENTIFIC VERSUS ORDINARY NOTIONS

One would therefore expect that any philosopher of mind or language would make it his or her business to understand the basic methodology and some of the results of this subject. But many philosophers of mind and language proceed in utter ignorance of the subject.

For example, it is or ought to be well known that linguists do not theorize using the ordinary notion of language in which German and Dutch are two different languages and Chinese is a single language. Language in this ordinary sense is a political or social notion. Speakers on both sides of the Dutch-German border understand each other quite well, although some are counted as speakers of German and some as speakers of Dutch, and both have trouble understanding other speakers of German. There are great differences between Chinese speakers, yet all are counted as speaking the same language, Chinese, whereas people in France and Portugal are not counted as speaking the same language, although their differences are much less than differences among Chinese. The social and political aspect of this conception of language is brought out in the quip that a language is a dialect with an army and a navy. So, some theorists at least take the unit of language to be a dialect or even an individual "idiolect"—the very particular dialect of a particular person, with no assumption that any two people have exactly the same idiolect. In fact, none of these notions—language, dialect, idiolect—plays any role in contemporary linguistic theory, which is concerned with characterizing the internal linguistic faculty.

Nevertheless, many philosophers persist in thinking that linguistics is or ought to be the study of language in this ordinary sense. They argue (or, rather, assert) that it is theoretically important to try to specify what it is to have a more or less complete grasp of a language in that sense and they suppose that people with an incomplete or erroneous understanding of aspects of a language do or should defer to others who have a more complete understanding These philosophers believe furthermore that principles of language have to do with social rules or conventions among people who speak the same language

This is true despite the fact that the principles of language of interest to contemporary linguistics are *never* explained as due to linguistic convention or social practice Contemporary linguistics supposes that the relevant principles are built into the language faculty and are therefore the same in all languages, up to parametric variation Language use is not a matter of rule following in any intelligible sense

It is true that a speaker may defer to one or another elite with respect to pronunciation or conditions of application for terms just as a person may defer to another in all sorts of other ways The notion of "misuse" of language is relevant to "the study of the sociology of group identification, authority structure, and the like" (71), but it is not taken seriously in linguistics

II MIND-BODY PROBLEM

Chomsky raises other related issues in these essays For example, he argues that it is a confusion to suppose that there is a "mind-body problem " Before Newtonian physics, the mind-body problem was the problem of giving a mechanical explanation of mind, the presupposition being that everything else could be given a mechanical explanation With the failure of that presupposition, the issue is unclear

Just as philosophers of language often fail to distinguish between ordinary notions and scientific notions of language, so philosophers of mind often fail to distinguish between ordinary notions like belief, sensation, and desire, on the one hand, and notions that appear in scientific theories Chomsky observes that there is no more reason to think that notions of desire or belief will play a role in scientific psychology than to think that the ordinary notion of language will play a role in linguistics or that ordinary notions like desk and chair will play a role in physics (This does not mean there are not desks, chairs, languages, desires, or beliefs Only that these notions are not suitable for scientific purposes) Similarly, there is no reason to assume, as many philosophers do, that mental "representations" appealed to in psychological or linguistic theories must represent things in the world

Contemporary philosophy of language is sometimes concerned with alleged relations between expressions and things, denotation, reference, where there is a certain amount of appeal to "intuitions "

Chomsky observes that we cannot have intuitions about these things deriving from our language faculty any more than we can have such intuitions about angular momentum

Furthermore, it is important to distinguish ethnoscience, which might be concerned with how people normally understand things, with physics or psychology or linguistics, which is concerned with what is actually the case Chomsky argues that philosophers are often best understood as doing ethnoscience, although without adhering to normal standards of empirical inquiry In any event, the study of the semantic resources of the language faculty is to be distinguished from the study of ordinary conceptions of meaning

III WORDS

Another theme in these essays has to do with the semantic representation of words in a natural language Chomsky argues that the underlying semantic representation of most words is quite complicated, often involving intricate and highly specialized perspectives involving human interests and concerns, providing various analytic connections in ways that could not be learned from scratch, so there is a "poverty of stimulus" argument here

Here are some examples If I say that I painted my house brown, you will understand me to mean I painted it brown on the outside That is the default or unmarked case But I can also say that I painted my house brown on the inside To climb a mountain is to go up, but you can also climb down the mountain If I am inside my house I cannot see it (except perhaps through a window if an exterior surface is then visible) I am not near my house if I am inside

Words offer conflicting perspectives The book I just took out of the library has a red cover, took four years to write, weighs two pounds, and has been translated into several languages I can paint the door brown and then walk through it After the bank lowered interest rates, it burned down and was rebuilt across the street The pronoun in this case can refer to a financial institution but cannot be used to refer back to a river bank, as in, 'After the bank lowered interest rates, there was a flood and the river overflowed it' This may relate to a worry some philosophers have about the idea that to know something you must believe it and it must be true What is known is a fact, what is believed is a proposition Some philosophers therefore object that what you know cannot be what you believe But perhaps that is like objecting that the baby cannot both finish the bottle and then break it

In this brief review I have been able to mention only a small part of the riches contained and I have not really been able adequately to defend Chomsky's complaints against philosophers of language and philosophers of mind, although I am convinced that he is right. In any event, this collection should be read and studied carefully by all philosophers of mind and language.

GILBERT HARMAN

Princeton University

Collected Papers. JOHN RAWLS. Cambridge: Harvard University Press, 1999. xxi + 656 p. Cloth $39.95.

Long before the publication of *A Theory of Justice,*[1] John Rawls deeply influenced moral and political philosophy through a series of articles that remain classics. Again in the years before the publication of *Political Liberalism,*[2] Rawls published important papers that foreshadowed the shift in perspective reflected in that second book. His more recent papers develop the idea of public reason central in *Political Liberalism* and lay the foundation for his *The Law of Peoples.*[3] Rawls had been reluctant to publish his collected papers, apparently because he saw them as "experimental," putting forth ideas to be reworked after further reflection and discussion. Until now, the papers have been scattered about in various journals and books, some of which are difficult to find. Now these significant papers have been collected in one volume, well edited by Samuel Freeman.

Collected Papers brings together nearly all of Rawls's published papers. These reflect the development of Rawls's thought from twenty years before *A Theory of Justice,* through his transition to *Political Liberalism,* and beyond. The collection includes twenty-five papers plus the Preface for the French Edition of *A Theory of Justice,* and an interview published in *Commonweal.* Freeman has provided a brief informative Preface and a thorough analytical index similar in style and scope to the indices that Rawls provided for *A Theory of Justice* and *Political Liberalism.*

The essays in *Collected Papers* fall into several groups. The first ten essays were written from 1951 to 1971, the period in which Rawls was constructing *A Theory of Justice.* The next five, written in the 1970s, are

---

[1] Cambridge: Harvard, 1971.

[2] New York: Columbia, 1993.

[3] Cambridge: Harvard, 1999.

0022-362X/01/9805/269–72

primarily extensions of *A Theory of Justice*, explaining some of its ideas more fully and responding to various objections. Next, six papers (and the Preface to French Edition of *A Theory of Justice*) represent Rawls's gradual transition to his political conception of justice as fairness, which culminates in *Political Liberalism*. The last four essays in *Collected Papers* take us beyond *Political Liberalism* in several ways. "The Idea of Public Reason Revisited" introduces modifications to be included in a later edition of *Political Liberalism*. "Themes in Kant's Moral Philosophy" is a sample of Rawls's influential work on the history of ethics, more of which is now available in a volume edited by Barbara Herman, *Lectures on the History of Moral Philosophy*.[4] The two remaining papers, "The Law of Peoples" and "Fifty Years After Hiroshima," are the basis for Rawls's somewhat fuller discussion of international justice, *The Law of Peoples*.

A few of Rawls's previous publications are not contained in *Collected Papers*. Some of these were omitted because they were early versions of papers later published in a revised and expanded form. This was the reason for excluding the initial versions of "Justice as Fairness" and "The Basic Structure as Subject."[5] Three other essays were not reprinted in *Collected Papers* because they are all republished in the paperback edition of *Political Liberalism*. These are "Reply to Habermas," "The Basic Liberties and Their Priority," and a later version of "The Basic Structure as Subject."[6] Finally, *Collected Papers* does not include a few short pieces written for special occasions. These are "Review of Toulmin," "Reply to Lyons and Teitelman," and "Roderick Firth: His Life and Work."[7]

Collected Papers should be welcomed by different sorts of readers for a variety of reasons. First, they provide a convenient packaging of essays with which, in addition to Rawls's books, any professional moral and political philosopher should be acquainted. Early papers, such as "Two Concepts of Rules" and "Outline of a Decision Procedure for Ethics," make significant proposals beyond what is developed in *A Theory of Justice*. It also may be instructive to see why Rawls abandoned ideas in earlier papers: for example, withdrawing a major argument about "real risks" in "Constitutional Liberty and the Concept of Justice" in response to an objection by Gilbert Harman. It remains an important, and much debated, question why Rawls developed a modified position in *Political Liberalism* and whether the changes were, as he thought, necessary improvements. The intervening essays help to clarify this. Some essays are important supplementary clarifications of points in the books, and the later essays lay out Rawls's interpretation of Kant's ethics and his treatment of international justice.

---

[4] Cambridge: Harvard, 2000.

[5] Respectively, in *Philosophical Review*, LXVII, 2 (April 1958): 164-94, and in Alvin I. Goldman and Jaegwon Kim, eds., *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt* (Boston: Reidel, 1978), pp. 47-71.

[6] Respectively, in this JOURNAL, XCII, 3 (March 1995): 132-80; in Sterling McMurrin, ed., *The Tanner Lectures on Human Values* (Salt Lake City: Utah UP, 1983), pp. 3-87; and in *American Philosophical Quarterly*, XIV (1977): 159-65.

[7] Respectively, in *Philosophical Review*, LX, 4 (October 1951): 572-80; in this JOURNAL, LXIX, 18 (October 5, 1972): 556-57; and in *Philosophy and Phenomenological Research*, LI, 1 (March 1991): 109-18.

Second, *Collected Papers* should interest students, teachers, and intelligent general readers who want to understand some main themes but are not prepared to read all of Rawls's work. The set of papers include some quite readable and challenging essays that are, compared to the books and other papers, not so heavily laden with detailed argument, replies to objections, and background presuppositions. "Justice as Fairness" and "Justice as Reciprocity" may be the best examples of this. Having the whole set together, of course, encourages further exploration.

Third, the volume is a valuable resource for scholars and students in any discipline that raises questions about the many topics that Rawls addresses. Thanks to the excellent index to *Collected Papers*, one can easily find where Rawls comments on various philosophical concepts, social issues, and other theories and authors, and his comments are invariably worth consulting.

Very briefly, some main lines of the moral and political philosophy expressed in *Collected Papers* as well as Rawls's previous books are these. The work starts from a conviction that we cannot satisfactorily address particular questions of justice without first determining the demands of justice on the basic structure of society. The initial step toward evaluating traditional principles for this purpose was to construct the idea of an "original position" which is arguably the appropriate standpoint for choosing among them. On this basis, Rawls then argues that two principles of justice he articulates are preferable to other traditional principles, such as utilitarianism and perfectionism. Rawls's principles insist on the priority of constitutional liberties, a strong requirement of "fair equality of opportunity," and a distributive principle that opposes inequalities in the basic structure that are not to the benefit of the least well off. Rawls attempts to confirm his results by applying the principles to show that they tend to match our "considered judgments" regarding particular matters. Then he considers how a "sense of justice" might develop to make a just society stable. Early papers, like his first book *A Theory of Justice*, do not sharply separate moral and political issues, and they suggest aspirations for developing a comprehensive, true moral theory. Later papers, like the later book *Political Liberalism,* consider the basic theory ("justice as fairness") only as a "political conception" on which, arguably, it is possible to develop an "overlapping consensus" among those who believe in diverse, but "reasonable," comprehensive moral and religious doctrines. Here, justification is sought by emphasizing potentially shared political values compatible with a wide diversity of views about morality and religion. Thus the principles of justice as fairness are presented without insisting on them as "true" or justifiable regardless of historical conditions. Rawls's papers as well as his books reflect his influential views on how moral and political theories can be constructed and justified. We are to aim, not for deductive proofs from self-evident premises, but rather for a "reflective equilibrium" in our thinking after we have surveyed all aspects of a theory and its alternatives in the light of whatever relevant facts and arguments we can think of.

Rawls's special vocabulary, with terms such as "reflective equilibrium," "original position," "primary goods," "the difference principle," and the like, has become part of the common currency of philosophical discourse. He has been often, and rightly, honored for his innovative, thorough, systematic, and deeply thoughtful treatment of the central problems of social justice. His work is also valued for the way it reconstructed and revitalized ideas prominent in the social-contract tradition and Kantian moral philosophy. His ideas have been controversial, but virtually all critics readily acknowledge the value of seriously engaging with his thought. But what has inspired generations of Rawls's students, more than any particular set of ideas, has been Rawls's example as a teacher, philosopher, and human being. The qualities they admire will be evident, I expect, to any reader of *Collected Papers*. For example, although the papers focus on abstract principles and arguments in a characteristically impersonal style, the reader can hardly fail to notice the author's humanity, fairness, intellectual integrity, and deep commitment to the pursuit of reasonable responses to social injustice.

THOMAS E. HILL, JR.

University of North Carolina/Chapel Hill

THE JOURNAL OF PHILOSOPHY

VOLUME XCVIII, NO 6, JUNE 2001

18171 2001

THE INTRANSITIVITY OF CAUSATION REVEALED IN EQUATIONS AND GRAPHS*

We live in exciting times By 'we' I mean philosophers studying the nature of causation The past decade or so has witnessed a flurry of philosophical activity aimed at cracking this nut, and, surprisingly, real progress has been made Two developments are especially worthy of note [1] First, there has been a resurgence of interest in the counterfactual theory of causation, given its best-known formulation by David Lewis [2] Second, there has been increasing philosophical interest in the techniques of causal modeling developed and employed within fields such as econometrics, epidemiology, and artificial intelligence [3] These two developments have been largely independent and addressed to different sorts of problems Work in the counterfactual tradition has been primarily concerned with issues involving "token" or "singular" causation, while work in the second tradition has tended to focus on issues concerning "type-level" or "general" causation

---

* For discussion and comments, thanks go to Martin Barrett, Paul Bartha, Ellery Eells, Clark Glymour, Alan Hájek, Dan Hausman, Mark Kalderon, Henry Kyburg, Michael McDermott, Laurie Paul, Agustin Rayo, Jonathan Schaffer, Elliott Sober, Jim Woodward, Stephen Yablo, and especially Judea Pearl

[1] I do not mean to disparage other important developments in the theory of causation, such as the development of a theory of causal processes in terms of conserved quantities—see Phil Dowe's *Physical Causation* (New York Cambridge, 2000)—but I shall not talk about them here

[2] "Causation," this JOURNAL, LXX, 17 (October 11, 1973) 556-67, reprinted in *Philosophical Papers, Volume II* (New York Oxford, 1986), pp 159-72 For examples of important recent work on counterfactual theories of causation, see the essays in this JOURNAL, XCVII, 4 (April 2000), and in John Collins, Ned Hall and L A Paul, eds *Causation and Counterfactuals* (Cambridge MIT, forthcoming)

[3] See, for example, the essays in Vaughan R McKim and Stephen P Turner, eds , *Causality in Crisis?* (Notre Dame University Press, 1997)

0022-362X/01/9806/273-99

Fortunately, these two developments are just beginning to merge: the computer scientists Judea Pearl and Joseph Halpern[4] have developed theories of token causation in terms of structural equations of the sort used in causal modeling. The account I shall present here bears a very strong resemblance to the theories developed by Pearl and Halpern. I shall not present their account explicitly and contrast it with mine, but I urge philosophers to read this work on their own.

As a case study, I shall explore the problem of the transitivity of causation from within the structural equations framework. Causation is transitive if and only if, whenever *a* causes *b* and *b* causes *c*, then *a* causes *c* as well. Many philosophers, notably Lewis, have claimed that causation is transitive. Others[5] have raised powerful objections to the transitivity of causation. I side with the objectors. This is not to say that I shall offer a conclusive refutation of the transitivity thesis: the transitivity fetishist may well be able to preserve her cherished principle by tweaking my account in various ways. But I shall show that there is no *independent* motivation for accepting the transitivity thesis: all of the benefits of maintaining transitivity can be had without paying the costs.

By working within the structural equations framework, I shall not be abandoning the traditional counterfactual framework. Rather, I shall present structural equations as tools for representing patterns of counterfactual dependence. These tools allow us to make explicit the structural differences between those cases which appear to instantiate the transitivity of causation, and those which appear to be counterexamples. The structural equations framework draws attention to a certain kind of counterfactual whose importance to the analysis of causation has hitherto been ignored by philosophers. I shall use this type of counterfactual to define the notion of an *active route* between two events. If it is possible to give a reductive analysis of counterfactuals in purely *acausal* terms, as Lewis[6] believes, then I offer a reductive analysis of active routes. In those cases which appear to be counterexamples to the transitivity of causation, the putative cause and effect fail to be connected by an active route.

---

[4] Pearl, *Causality* (New York: Cambridge, 2000), chapter 10, Halpern and Pearl, "Causes and Explanations: A Structural-model Approach," Technical report R-266, Cognitive Systems Laboratory, University of California/Los Angeles, 2000.

[5] See especially Michael McDermott, "Redundant Causation," *British Journal for the Philosophy of Science*, XL (1995): 523-44.

[6] See, for example, "Counterfactual Dependence and Time's Arrow," *Noûs*, XIII (1997): 455-76, reprinted in *Philosophical Papers, Volume II*, pp. 32-52, and *Philosophical Papers, Volume II*, introduction.

In focusing on the problem of transitivity, I shall have little or nothing to say about a variety of other problems in the theory of causation. I shall assume determinism, and hence leave unsolved important problems involving indeterministic causation. I shall discuss only briefly the problems of symmetric overdetermination, late preemption, and "trumping." My strategy will be to take Lewis's account of causation as a foil,[7] argue that my account does a better job on the issue of transitivity, and say just enough about these further issues to suggest that my account fares no worse in these arenas.

I. TRANSITIVITY: ANATOMY OF A PUZZLE

Our point of departure will be Lewis's original counterfactual theory of causation. Let *c* and *e* be distinct events that both occurred. Then *e counterfactually depends upon c* if and only if, if *c* had not occurred, *e* would not have occurred. Lewis takes counterfactual dependence to be sufficient but not necessary for causation: causation is defined as the *ancestral* of counterfactual dependence, rendering causation transitive by definition.

Lewis[8] offers a detailed account of what must be "held fixed" when evaluating the relevant counterfactuals. For our purposes, it is enough to point to two features that these counterfactuals must have. First, the counterfactuals must not *backtrack*. If *a* caused *c* (but not vice versa), then a counterfactual of the form 'If *c* had not occurred, then ...' must hold *a* fixed. If not, the theory would incorrectly rule that *c* causes *a*. Second, the counterfactuals must "foretrack": if *c* causes *e*, we do not want to hold *e* fixed when evaluating the counterfactual 'If *c* had not occurred, then ...' If we do, the consequent of the conditional will obviously not be '*e* would not have occurred'.

Note that these are restrictions on what is to be *tacitly* held fixed when entertaining counterfactual antecedents; they are not restrictions on the antecedents that we are permitted to entertain. The necessity of foretracking from *c* to *e* does not prevent us from entertaining a counterfactual of the form 'If *c* had not occurred, but *e* had occurred anyway, then ...'[9] Let us call a counterfactual of this sort an *explicitly nonforetracking* or ENF counterfactual. In the sequel, I shall argue that ENF counterfactuals should play a central role in the analysis of causation.

---

[7] I do not deny that other counterfactual theories of causation would make worthy foils. Space limitations prohibit detailed comparison with all such rivals.

[8] "Counterfactual Dependence and Time's Arrow."

[9] Lewis himself makes this point in "Finkish Dispositions," *Philosophical Quarterly*, XLVII (1997): 143-58, here p. 150.

Why does Lewis define causation as the *ancestral* of counterfactual dependence? There is certainly a strong pretheoretic intuition that causation is transitive, but this can be explained without building transitivity into the analysis Let us call a case of causation *ordinary* if it has this structure *e* depends counterfactually upon *d*, which in turn depends counterfactually upon *c*, and *e* also depends counterfactually upon *c* *Most* cases of causation are ordinary, and in such cases we can explain why *c* counts as a cause of *e* just by identifying causation with counterfactual dependence

Trouble arises in "extraordinary" cases In many such cases, we judge that there is causation without counterfactual dependence Here is a standard example

> "Backup" an assassin-in-training is on his first mission Trainee is an excellent shot if he shoots his gun, the bullet will fell Victim Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger If Trainee does not shoot, Supervisor will shoot Victim herself In fact, Trainee performs admirably, firing his gun and killing Victim

In this case, it seems that Trainee's shot caused the death of Victim, even though Victim's death does not counterfactually depend upon Trainee's shot This is a case of *preemption* by shooting at Victim, Trainee preempted a process that would itself have resulted in Victim's death The standard solution is to invoke the transitivity of causation Trainee's shot is a cause of Victim's death because there is a chain of counterfactual dependence running from the former to the latter Consider a further event (call it '*b*') that was not mentioned explicitly in the foregoing description the presence of a bullet en route from Trainee to Victim Had Trainee not shot, *b* would not have occurred, and if *b* had not occurred, Victim would not have died Note the importance of the "no backtracking" rule in the second counterfactual if *b* had not occurred, Trainee would have shot anyway, so Supervisor would not have shot

Other examples suggest, however, that causation is not transitive in general

> "Boulder" a boulder is dislodged, and begins rolling ominously toward Hiker Before it reaches him, Hiker sees the boulder and ducks The boulder sails harmlessly over his head with nary a centimeter to spare Hiker survives his ordeal [10]

---

[10] This example is from an early draft of Hall's "Two Concepts of Causation," forthcoming in Collins et alia

> "Dog Bite" Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb Shortly before noon, he is bitten by a dog on his right hand Unable to use his right hand, he pushes the detonator with his left hand at noon The bomb duly explodes [11]

In each case, we have a chain of counterfactual dependence Hiker would not have ducked if the boulder had not fallen, and given the trajectory of the boulder, he would not have survived if he had not ducked The dog bite caused Terrorist to push the detonator with his left hand, and his pushing the detonator with his left hand caused the bomb to explode But the fall of the boulder did not cause Hiker to survive and the dog bite did not cause the explosion—these are the verdicts of common sense

We should be particularly troubled that we judge there to be a causal relationship in that case where the chain of counterfactual dependence is hardest to see In "Backup," we must find an intermediate event that is not made salient in the presentation of the example, namely, *b*, and imagine it away while holding Trainee's shot fixed We are to imagine the bullet vanishing into thin air or some such This is no homey piece of counterfactual reasoning, but requires the efforts of a trained philosopher In "Boulder" and "Dog Bite," by contrast, it is easy to see the needed intermediate events Hiker's ducking and Terrorist's pushing the detonator with his left hand (respectively) These two events were specified as parameters of the examples Moreover, the necessary counterfactual reasoning is not particularly straining it is easy to imagine, for example, what would have happened if Hiker had not ducked, even holding fixed the boulder's fall The defender of transitivity must maintain that we have inconsistently acute intuitive powers we see causal relations that are underwritten by obscure chains of counterfactual dependence, and yet we are blind to causal relations that are underwritten by obvious chains of counterfactual dependence

I shall argue that we can accept "Dog Bite" and "Boulder" as counterexamples and provide an alternative account of "Backup " Trainee's shot caused Victim's death, not because there is a chain of counterfactual dependence, but because there is an *active route* between Trainee's shot and Victim's death The existence of this route is revealed by an ENF counterfactual if Trainee had not shot, and Supervisor still did not shoot, then Victim would not have died There are no comparable ENF counterfactuals in "Boulder" and "Dog Bite "

---

[11] From McDermott, "Redundant Causation "

II FURTHER PROBLEMS

While my primary focus will be on the problem of transitivity, I shall briefly mention a few further problems with Lewis's theory along with his recent attempt[12] to solve some of them The first sort of problem involves cases of *symmetric overdetermination,* in which two events have an equal claim to be causes of a third event, which does not depend counterfactually upon either of them Lewis[13] claims that he has no clear intuitions about such cases, and thus dismisses them as being of no diagnostic value

A second type of problem involves cases of preemption that have different structures from "Backup " Lewis[14] uses the term *early cutting* to describe this structure In addition, he recognizes cases of *trumping*[15] and *late cutting* Here is an example of the latter

> Billy and Suzy both throw rocks at a bottle Suzy's throw gets there first, shattering the bottle Billy's throw arrives at the scene a split second later, encountering nothing but air where the bottle used to be [16]

Suzy's throw is clearly a cause of the bottle's shattering, Billy's not The shattering of the bottle does not counterfactually depend upon Suzy's throw, and there is no chain of counterfactual dependence from Suzy's throw to the bottle's shattering [17]

Lewis responds to these problems by modifying his theory of causation He first defines the notion of *influence*

> Where c and e are distinct actual events, c influences e if and only if there is a substantial range c_1, c_2, of different not-too-distant alterations of c (including the actual alteration of c) and there is a range e_1, e_2, of alterations of e, at least some of which differ, such that if c_1 had occurred, e_1 would have occurred, and if c_2 had occurred, e_2 would have occurred, and so on [18]

(An alteration c' of c is a fine-grained event that is similar to c, but possibly different in matters of detail) Suzy's throw influences the shattering of the bottle had Suzy thrown slightly earlier, or aimed at a slightly different point, the bottle would have shattered slightly earlier, or in a slightly different way Lewis then defines causation as the ancestral of influence, citing the problem of early cutting preemption as his motivation for doing so [19]

---

[12] "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000) 182-97

[13] Both in "Causation" and in "Causation as Influence "

[14] "Causation as Influence "

[15] For a detailed discussion of trumping, see Schaffer, "Trumping Preemption," this JOURNAL, XCVII, 4 (April 2000) 165-81

[16] This example appears in "Causation as Influence," but has been in the lore for some time

[17] For a defense of the latter claim, see "Postscript E" to "Causation" in Lewis's *Philosophical Papers, Volume II,* pp 193-213

[18] "Causation as Influence," p 190

I conclude this section by pointing to an interesting feature of Lewis's new account it is infected with context sensitivity In order for *c* to influence *e*, there must be true counterfactuals involving a *substantial range* of *not-too-distant* alterations of *c* It may be an objective matter whether a set of counterfactuals of the form 'If c_i had occurred, e_i would have occurred' are true, but there is a further question about whether we should describe this pattern by saying '*c* causes *e*' The answer to this question will depend, in part, upon which unactualized possibilities we consider "too distant" to take seriously This point is underscored in Lewis's[20] brief discussion of *preemptive prevention* The account developed below will agree with Lewis's new account on this point

III EQUATIONS AND GRAPHS FOR DUMMIES

The use of directed graphs to represent systems of causal relationships dates back at least to the work of Sewall Wright[21] in the early 1920s, the use of structural equations, pioneered by R Frisch,[22] T Haavelmo[23] and others in the 1930s and 1940s is not much younger Pearl's *Causality* is representative of the current state of the art [24]

A system of structural equations is a sequence of equations $\mathscr{E}$ relating the values of variables belonging to some set $\mathscr{V}$ An ordered pair $\langle \mathscr{V}, \mathscr{E} \rangle$ will be called a *causal model* In the simplest case, a variable will have only two values, and will represent the occurrence or nonoccurrence of a certain event For example, in representing the causal relations in "Backup," we shall use a variable T, which takes the value 1 when Trainee shoots, and takes the value 0 when he does not As a general convention, variables will be represented by italicized capital letters When a binary variable E is used to represent the occurrence or nonoccurrence of an event e, $E = 1$ will represent the occurrence of e, and $E = 0$ the nonoccurrence of e But variables need not be binary For example, the values of a variable could represent the mass of some object, or they could represent various alterations of a particular event $\mathscr{V}$ contains both *exogenous* and *endogenous* variables

---

[19] See especially "Causation as Influence," p 194 Note that in Lewis's new account, transitivity would no longer be needed to handle "Backup" the precise time and manner of Victim's death does indeed depend upon the precise time and manner of Trainee's shot Lewis maintains that transitivity will still be needed to handle fancier examples of early cutting I shall continue to use "Backup" in unmodified form, this will not change the fundamental structure of our central problem

[20] "Causation as Influence," p 197 For more detailed discussion, see Collins, "Preemptive Prevention," this JOURNAL, XCVII, 4 (April 2000) 223-34

[21] "Correlation and Causation," *Journal of Agricultural Research*, XX (1921) 557-85

[22] "Statistical versus Theoretical Relations in Economic Macrodynamics," League of Nations Memorandum, 1938, reprinted as "Autonomy of Economic Relations" in D Hendry and M Morgan, eds , *The Foundations of Econometric Analysis* (New York Cambridge, 1995), pp 407-19

[23] "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, XI (1943) 1-12, reprinted in Hendry and Morgan, pp 454-63, "The Probability Approach in Econometrics," Supplement to *Econometrica*, XII, reprinted, in part, in Hendry and Morgan, pp 440-53, 477-90

[24] Philosophers will probably be most familiar with the work of Peter Spirtes, Clark Glymour, and Richard Scheines, especially their *Causation, Prediction, and Search* (New York Springer, 1993) This work is very much in the same vein, although I shall not be using their probabilistic framework here

Each equation in $\mathscr{E}$ expresses the value of one variable, appearing on the left-hand side of the equation Each variable appears on the left-hand side of exactly one such equation $\mathscr{E}$ is comprised of two subsets, $\mathscr{E}_I$ and $\mathscr{E}_{II}$ The former contains equations with exogenous variables on the left-hand side, while the latter contains equations with endogenous variables on the left hand side Equations in $\mathscr{E}_I$ all take the simple form $X = x$ they simply state the actual value of the variable in question Equations in $\mathscr{E}_{II}$ express the value of the endogenous variable as a function of the values of other variables in the set $\mathscr{V}$

$$(*)\ Z = f_Z(X, Y, \ , W)$$

The syntax of such an equation is richer than that of ordinary mathematical equations In particular, structural equations are not symmetric (*) is not equivalent to $f_Z(X, Y, \ , W) = Z$. (In structural equations, side matters) This is because structural equations encode counterfactuals For example, (*) encodes a set of counterfactuals of the following form

> If it were the case that $X = x$, $Y = y$, , $W = w$, then it would be the case that $Z = f_Z(x, y, \ , w)$

These counterfactuals are to be understood along the lines discussed in section I above, in particular, they do not backtrack

Equations in $\mathscr{E}_{II}$ must always be written in minimal form if for all x, x', y, z, , w, $f_Z(x, y, \ , w) = f_Z(x', y, \ , w)$, then the value of Z does not depend upon the value of X at all, and the structural equation for Z must be rewritten $Z = f_Z(Y, \ , W)$ As written, then, the equation (*) says that the value of the variable Z depends counterfactually upon the values of the variables X, Y, W By the same token, equations in $\mathscr{E}_{II}$ must always include as arguments any variables in $\mathscr{V}$ upon which Z counterfactually depends, given the values of the other variables If, for some x, x', y, z, , w, $f_Z(x, y, \ , w) \neq f_Z(x', y, \ , w)$, then the value of Z does depend upon the value of X, and $Z = f_Z(Y, \ , W)$ is not in $\mathcal{E}$ The correct equation for Z can be arrived at by expressing the value of Z as a function of *all* other variables in $\mathcal{V}$, and then eliminating those variables whose values are redundant given every assignment of values to the other variables

It will sometimes be helpful to use symbols familiar from sentential logic to represent relations between variables The symbols $\neg$, $\vee$, and $\wedge$ will represent the following mathematical functions $\neg X \equiv 1 - X$, $X \vee Y \equiv \max\{X, Y\}$, $X \wedge Y \equiv \min\{X, Y\}$ When the variables are binary, these functions behave much as the corresponding connectives do in sentential logic For example, if $Z = X \vee Y$, then Z will take the value 1 if and only if either X or Y takes the value 1

If the variable X figures as an argument on the right-hand side of the structural equation for the endogenous variable Z, then X is a *parent* of Z Exogenous variables have no parents within a model, while endogenous variables do

A system of structural equations can be given an elegant graphical representation The variables in $\mathcal{V}$ form the nodes of a graph The nodes are connected by directed edges or "arrows" according to the following rule an arrow is drawn from X to Z if and only if X is a parent of Z A *directed path* from variable X to variable Z is a sequence of arrows lined up "tip-to-tail" connecting X with Z A variable is exogenous if there is no arrow directed into it

If the structural equations can be ordered so that no variable appears on the left-hand side after having appeared on the right-hand side, then the system of equations is *acyclic* Equivalently, a system of structural equations is acyclic if no directed path in the corresponding graph runs from a given node back into itself Intuitively, an acyclic system of equations represents a causal structure in which there are no causal loops All of the systems considered here will be acyclic If a system of equations is acyclic, then the system of equations has a unique solution That is, the values of the exogenous variables together with the other structural equations entail a unique value for every variable

I shall illustrate structural equations and graphical representations using "Backup " The causal graph is depicted in figure 1 The variables are to be interpreted as follows $T = 1$ corresponds to Trainee's shooting, $T = 0$ to his refraining, $S = 0$ or 1 depending upon whether Supervisor shoots, $B = 0$ or 1 depending upon the presence of a bullet in flight at some point along the line from Trainee to Victim, and $V = 0$ or 1 according to whether Victim dies The set of structural equations is

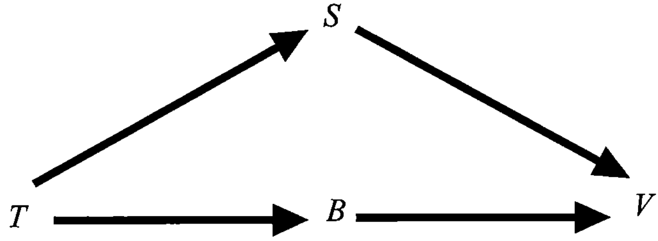


Figure 1

$$\mathscr{SE}: T = 1,\ S = \neg T,\ B = T,\ V = B \vee S$$

T is an exogenous variable. The equation, $V = B \vee S$ encodes the following counterfactuals: if either B or S were to take the value 1—if either *b* or Supervisor's shot were to occur—then Victim would die; if B and S were both to take the value 0, then Victim would not die. $\mathscr{SE}$ has the following unique solution:

$$T = 1,\ S = 0,\ B = 1,\ V = 1$$

That is: Trainee actually shot, Supervisor did not shoot, *b* occurred, and Victim died.

It has become common in the philosophical literature to represent causal relationships in terms of *neuron diagrams*. A circle is used to represent an event which may occur ("fire") or not. A shaded circle represents an event that did occur, while a hollow circle represents an event that did not occur. Neurons may be connected in one of two ways: by a stimulatory connection, represented by an arrow, or by an inhibitory connection, represented by an arrow with a backward head. A stimulatory connection between two events indicates that, if the first occurs, it will cause the second to occur; an inhibitory connection indicates that, if the first occurs, it will prevent the second from occurring (inhibitory connections override stimulatory ones). Causal graphs work similarly, but with several important differences. First, each vertex represents a *variable*. A variable may be binary, taking values of 1 or 0 according to whether some event occurs or not, but it may also be multivalued. Second, the causal graph does not itself specify the actual value of the variable; that information is contained in the accompanying set of structural equations. Third, an arrow from one variable to another says nothing about the *kind* of connection that exists between them—once one allows nonbinary variables, the possible connections cannot be classified neatly into stimulatory and inhibitory[25] I have argued elsewhere[26] that attempts to shoe-horn all causal relationships into these two types lead to a number of pseudo-problems in the theory of causation Instead, the nature of the connection between variables is represented in the corresponding system of structural equations

Each equation in $\mathscr{E}_{\mathscr{V}}$ encodes counterfactual information Note, however, that the equations in $\mathscr{E}_{\mathscr{V}}$ do not directly represent *all* counterfactuals that are true of the system Rather, $\mathscr{E}$ is a set of *fundamental* equations from which all other counterfactuals may be derived In general, in order to evaluate the counterfactual 'If it were the case that $X = x$, $Y = y$, , $W = w$, then ', we replace the equation for each of these variables with the identity stipulated, for example, we replace the equation for X with $X = x$ The equations for the other variables remain unchanged In effect, this creates a new set of structural equations in which X, Y, , W are exogenous variables Graphically, the arrows directed into these variables are removed, while all other arrows remain intact Instead of these variables having their values causally determined in the normal way, they are "miraculously" set to the new hypothetical values The values of the remaining variables can then be computed The result may be thought of as characterizing the "closest possible world(s)" where $X = x$, $Y = y$, , $W = w$, are all true A variable Z depends counterfactually upon a variable X in a system of structural equations if and only if in the actual solution, $X = x$, $Z = z$ and there exist $x' \neq x$ and $z' \neq z$ such that the result of replacing the equation for X with $X = x'$ yields $Z = z'$ This says that there is some possible value of X such that if X had taken that value, then the value of Z would have been different

The standard treatment of "Backup" requires the counterfactual 'If the bullet had not been in flight from Trainee to Victim, then Victim would not have died' Modifying $\mathscr{SE}$ by setting $B = 0$, we get the following modified set of structural equations

$$\mathscr{SE}' \quad T = 1,\ S = \neg T,\ B = 0,\ V = B \vee S$$

---

[25] Indeed, even with binary variables, this problem can arise as soon as you have two "neurons," each connected to a third Suppose that the third neuron fires if exactly one of the others does, then the connections are neither purely stimulatory nor purely inhibitory

[26] "A Generalized Probabilistic Theory of Causal Relevance," *Synthese*, XCVII (1993) 335-64, "The Mishap at Reichenbach Fall Singular vs General Causation," *Philosophical Studies*, LXXVIII (1995) 257-91, "Farewell to Binary Causation," *Canadian Journal of Philosophy*, XXVI (1996) 267-82, "The Role of Contrast in Causal and Explanatory Claims," *Synthese*, CVII (1996) 395-419

In the graphical representation of $\mathcal{SE}'$, the arrow from T to B is removed from figure 1 Solving, we have

$$T = 1, S = 0, B = 0, V = 0$$

So if the bullet had not been in flight, Trainee still would have shot, Supervisor would not have fired, and Victim would not have died—just as required by the standard solution This example illustrates how counterfactuals do not backtrack within the structural equations approach

A system of structural equations is an elegant means for representing a whole family of counterfactuals of just the sort that Lewis's counterfactual theory of causation depends upon The correctness of a set of structural equations, and of the corresponding graph, depends upon the truth of these counterfactuals If, as Lewis believes, the truth values of counterfactuals supervene upon noncausal facts, then the correctness of a set of structural equations does as well

IV CAUSAL ROUTES

Consider the causal structure depicted in figure 2 What does the arrow from X to Z mean in such a diagram? This is a fundamental question, whose answer serves to highlight an important difference between the structural equations approach to causation, and the counterfactual approach more familiar to philosophers

Note first the arrows from X to Y and from Y to Z These indicate that the value of Z depends counterfactually upon the value of Y, which in turn depends upon the value of X So far, so good When these relationships hold, the traditional counterfactual approach to causation leads us to ask two further questions Does the value of Z depend counterfactually upon the value of X? If not, is the value of X nonetheless a cause of the value of Z? But the arrow from X to Z in figure 2 does not correspond to an affirmative answer to either of these questions Rather, the arrow from X to Z means that the structural equation for Z is of the form $Z = f_Z(X, Y)$, where X is an

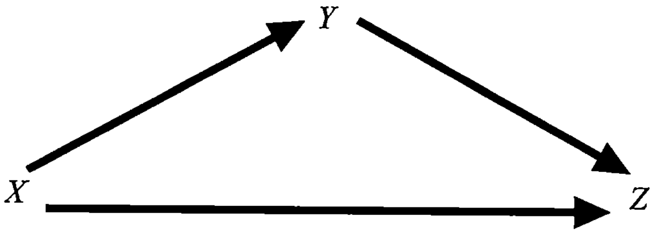


Figure 2

essential argument there exist x, x', y, such that $f_Z(x, y) \neq f_Z(x', y)$ An arrow from X to Z thus means that the value of Z can depend counterfactually upon the value of X, *even holding fixed the value of* Y The natural causal interpretation of this counterfactual is that the value of X can have an effect on the value of Z over and above the effect it has in virtue of causing the value of Y There are two routes whereby X influences Z, one which runs through Y, and one direct route which bypasses Y The overall effect of X on Z will depend upon both of these routes

Pay attention to the counterfactual that reveals the direct route from X to Z in figure 2 it requires that we hypothetically change the value of X, while holding the value of Y fixed, and evaluate whether Z changes in response This is an ENF counterfactual it holds fixed Y, an effect of X, in supposing X to take on a different value, it does not allow the counterfactual to foretrack from X to Y

"Traditional" counterfactual approaches to causation—those in the tradition of Lewis—do not employ ENF counterfactuals They employ only counterfactuals with simple antecedents, antecedents that make stipulations about only one event As a result, they are unable to detect the direct route from X to Z in figure 2 Indeed, despite the extensive use of neuron diagrams by philosophers working within the counterfactual tradition, neither I nor anyone I have polled has seen a diagram in which three neurons are configured in the triangular pattern shown in figure 2[1]

Whether a route is direct or not is relative to the variable set $\mathscr{V}$ It may be, for instance, that relative to the variable set $\mathscr{V}$, this route is direct, while in the richer variable set $\mathscr{V}'$, the route from X to Z that bypasses Y is mediated by some further variable W in $\mathscr{V}'\backslash\mathscr{V}$ If it is possible to interpolate a variable W along the direct route from X to Z, the traditional counterfactual approach *is* able to detect the existence of two routes from X to Z one via Y and the other via W This ability is attested to by the abundance of such "diamond-shaped" configurations in neuron diagrams Since the traditional counterfactual approach is able to distinguish causal routes *only* by interpolating variables, it is hardly surprising that the standard solution to cases of preemption like "Backup" is to interpolate and find a chain of counterfactual dependence But even when interpolation is possible and ENF counterfactuals are not needed to detect the *existence* of distinct causal routes, ENF counterfactuals provide crucial information about the *nature* of the distinct causal routes, and this is the key to solving our problem

Returning to figure 2, note that the arrow from X to Z implies only that there is *some* value y of Y such that $f_Z(x, y) \neq f_Z(x', y)$—such that

Z depends counterfactually upon X while fixing Y at y It may be, however, that given the *actual* value y of Y, Z does not depend upon X In such a case, we shall say that the direct route from X to Z is *inactive* Suppose, for example, that all three variables are binary and that $Z = X \wedge Y$ Then, if Y takes the value 1, the direct route from X to Z will be *active* whether $Z = 0$ or 1 will depend upon whether $X = 0$ or 1 (while holding $Y = 1$ fixed) In such a case, the value of X does in fact play a role in determining the value of Z, over and above the role it plays by determining Y On the other hand, if $Y = 0$, then the direct route from X to Z is inactive

Let us formulate this distinction between active and inactive routes more precisely, at the same time generalizing to include cases in which there are more than two routes between two variables, and where the route being evaluated is not direct Let $\mathscr{E}$ be a system of structural equations on the variables in set $\mathscr{V}$ A *route* between two variables X and Z in $\mathscr{V}$ is an ordered sequence of variables $\langle X, Y_1, \ , Y_n, Z\rangle$ such that each variable in the sequence is in $\mathscr{V}$, and is a parent of its successor in the sequence Graphically, a route between X and Z is a directed path from X to Z A variable Y, distinct from both X and Z, is *intermediate between X and Z* if and only if it belongs to some route between X and Z Then

> "Act" The route $\langle X, Y_1, \ , Y_n, Z\rangle$ is *active* in the causal model $\langle \mathscr{V}, \mathscr{E}\rangle$ if and only if Z depends counterfactually upon X within the new system of equations $\mathscr{E}'$ constructed from $\mathscr{E}$ as follows for all $Y \in \mathscr{V}$, if Y is intermediate between X and Z, but does not belong to the route $\langle X, Y_1, \ , Y_n, Z\rangle$, then replace the equation for Y with a new equation that sets Y equal to its actual value in $\mathscr{E}$ (If there are no intermediate variables that do not belong to this route, then $\mathscr{E}'$ is just $\mathscr{E}$)

The activity of a route is entailed by the truth of a certain kind of ENF counterfactual That is, the route is active if there is a true counterfactual of the form if the value of X had been x', and the value of variables that lie along *other* routes from X to Z were held fixed, then the value of Z would have been different By holding fixed intermediates along other routes, any influence of X on Z along those other routes is eliminated—hypothetical changes in X are not allowed to foretrack along those routes The relevant counterfactual thus isolates the influence of X on Z along the route in question

In analogy with Lewis's definition of influence, we could emend "Act" to require that there be "substantial" counterfactual dependence of Z on X in the new system of equations This would allow us to deny that a route is active, for example, if X is multivalued and very few values of X lead to a different value of Z, or if Z is multivalued, and changes in the value of X lead only to minimal changes in the value of Z Such an emendation would sacrifice precision, but better fit the coarse grain of our ordinary causal judgments I shall leave this choice to the reader, the emendation will not be necessary for any of our central examples

My central proposal is

> Let c and e be distinct occurrent events, and let X and Z be variables such that the values of X and Z represent alterations of c and e respectively Then c is a cause of e if and only if there is an active causal route from X to Z in an appropriate causal model $\langle \mathscr{V}, \mathscr{E} \rangle$[27]

What makes a causal model $\langle \mathscr{V}, \mathscr{E} \rangle$ appropriate? There are at least three requirements The first two are objective the equations in $\mathscr{E}$ must entail no false counterfactuals, and they must not represent counterfactual dependence relations between events that are not distinct The third component is pragmatic $\mathscr{V}$ should not contain variables whose values correspond to possibilities that we consider to be too remote I shall discuss these issues at greater length in sections VII and VIII below For now, note that these restrictions are present in Lewis's recent account and hence are not peculiar to the present account

In sections V, VII, and VIII below, I shall return to our three central examples I shall show that in "Backup," there is a causally active route from Trainee's shot to Victim's death This gives us the result that Trainee's shot causes Victim's death without the need to invoke transitivity This opens to door to the possibility of accepting "Boulder" and "Dog Bite" as counterexamples to the transitivity of causation, which I shall do I shall show that there is no active causal route from the dog bite to the explosion, or from the boulder's fall to Hiker's survival

V "BACKUP" REVISITED

Let us represent the causal structure of "Backup," this time omitting the esoteric variable B I claim as a virtue for my account that consideration of this variable is unnecessary (albeit harmless) The causal graph is depicted in figure 3, and the set of structural equations is

---

[27] In "A Tale of Two Effects" (forthcoming), I take a slightly different line Sometimes, our judgments of causation attach to individual active routes, at other times, we focus on the net effect of one variable on another along all routes taken together (and indeed there may even be cases where we attend to the effect along some intermediate number of routes) Nonetheless, it is the notion of dependence along a route captured by "Act" that seems to be at work in our judgments of token causation in the cases at issue here

$\mathscr{SE}$ $T = 1,\ S = \neg T,\ V = T \vee S$

The interpretation of the variables is the same as that given in section III above Figure 3 clearly shows two distinct routes from T to V, and it follows from the equations $\mathscr{SE}$ that these two routes "cancel" V does not depend counterfactually upon T

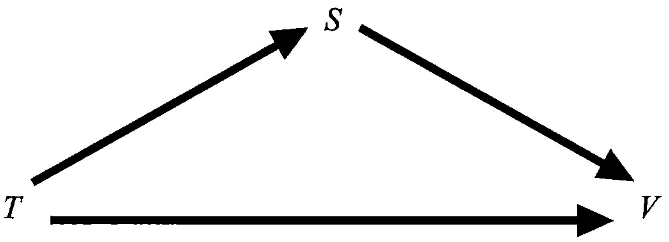


Figure 3

In order to see that the route $\langle T, V\rangle$ is active, we must show that V depends counterfactually upon T while holding the value of S fixed In fact, $T = 1$, $S = 0$, and $V = 1$ So we must show that if T had been 0, and S had also been 0, then V would have been 0 We must solve for the new structural equations

$\mathscr{SE}'$ $T = 0,\ S = 0,\ V = T \vee S$

It is easy to see that within this new system of equations, V must equal 0

We have just demonstrated the truth of the following ENF counterfactual if Trainee had not shot, and Supervisor had not shot either, then Victim would not have died The truth of this counterfactual is straightforward, and could have been evaluated intuitively without the use of formal apparatus It is in virtue of the truth of this counterfactual that Trainee's shot is a cause of Victim's death No wonder, then, that we have such an easy time judging that Trainee's shot is a cause of Victim's death, this treatment of "Backup" is much simpler than that required by the standard solution And for those[28] who worry about such other-worldly possibilities, the present account (unlike Lewis's) would work equally well if the assassins' guns did not kill by firing bullets, but rather by some sort of unmediated action at a distance

---

[28] For example, McDermott ("Redundant Causation") and Schaffer ("Trumping Preemption") discuss cases involving magic spells that act at a spatiotemporal distance

Note that in order to know that S must be held fixed, we do *not* need to know that Trainee's shot *caused* Supervisor not to shoot We only need to know that the pattern of counterfactual dependence is as described by $\mathcal{BU}$ Thus the need to hold fixed an event that is intermediate between Trainee's shot and Victim's death does not undermine the reduction of causation to patterns of counterfactual dependence

VI FURTHER PROBLEMS REVISITED

Before examining the two counterexamples to transitivity, I wish to return briefly to the problems discussed in section ii It should be clear that Lewis's solution to the problems of late cutting and trumping preemption are readily adapted to the present framework We could let the variable S take on different values representing different alterations of Suzy's throw, likewise, we could let values of the variable B represent differences in the time and manner of the bottle's shattering Then the value of B will depend counterfactually upon the value of S, and the route from S to B will be active

If the values of two variables X and Y symmetrically overdetermine the value of Z (for example, if all the variables are binary and $Z = X \vee Y$) then "Act" does not indicate an active route from either X or Y to Z To the extent that this is a shortcoming of "Act," it is a shortcoming of Lewis's theory as well Thus, these three further problems do not accord Lewis's theory any advantage over that presented here

It is worth briefly noting that the structural equations approach suggests novel solutions to the problems of late cutting preemption and symmetric overdetermination Detailed accounts are provided by Halpern and Pearl [29] Their formulations are slightly different from mine, but their proposals are readily adapted to the account sketched above Consider first the case of late cutting preemption Even without considering fine differences in the way Suzy throws her rock, or in the way the bottle shatters, it can be shown that there is an active route from Suzy's throw to the bottle's shattering This is revealed by the following ENF counterfactual *given* that Billy's rock did not hit the bottle, if Suzy had not thrown, the bottle would have remained intact throughout the incident

The problem of symmetric overdetermination can be handled by weakening "Act " For reasons of space, I will provide only a compact

---

[29] In Pearl, *Causality*, and Halpern and Pearl, "Causes and Explanations "

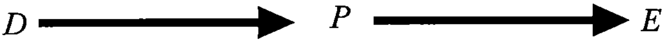


Figure 4

presentation of the technical details Let $\langle X, Y_1, \; , Y_n, Z\rangle$ be a route between the variables X and Z Suppose that the actual values of the variables on the route (other than X) are $Y_1 = y_1, \; , Y_n = y_n, Z = z$ Let $\{W_1, \; , W_m\}$ be a set of variables in $\mathcal{V}$ that does not belong to the route $\langle X, Y_1, \; , Y_n, Z\rangle$, these variables need not be intermediate between X and Z Let $w_1, \; , w_m$ be possible values of the variables $W_1, \; , W_m$, respectively The values $w_1, \; , w_m$ lie in the *redundancy range* of the variables $W_1, \; , W_m$ for the route in question if the following counterfactual is true if it were the case that $W_1 = w_1, \; , W_m = w_m$, then it would be the case that $Y_1 = y_1, \; , Y_n = y_n, Z = z$ In other words, we can set the values of the variables W_i to w_i without disturbing the variables (other than possibly X) on the route $\langle X, Y_1, \; , Y_n, Z\rangle$ It should be obvious that the actual values of $W_1, \; , W_m$ will lie in the redundancy range, but there may be nonactual values of these variables that do as well Indeed, cases of symmetric overdetermination arise precisely because nonactual values of these variables lie in the redundancy range We may now provide our generalization of "Act"

> *WA* The route $\langle X, Y_1, \; , Y_n, Z\rangle$ is *weakly active* relative to $\langle \mathcal{V}, \mathcal{E}\rangle$ if and only if there exists a set (possibly empty) of variables $\{W_1, \; , W_m\}$ in $\mathcal{V} \setminus \{X, Y_1, \; , Y_n, Z\}$, and values $w_1, \; , w_m$ that lie within the redundancy range of these variables for this route, such that Z depends counterfactually upon X within the new system of equations $\mathcal{E}'$ constructed from $\mathcal{E}$ as follows for each W_i, replace the equation for W_i with a new equation that sets W_i equal to w_i [30]

"Act" characterizes a special case where each W_i lies on a route from X to Z and is set to its actual value The reader may verify that *WA* is met in cases of symmetric overdetermination I fully grant that *WA* is less intuitive and less well-motivated than "Act" Since none of our central examples involves symmetric overdetermination, I shall continue to use "Act" in what follows [31]

VII "DOG BITE" REVISITED

"Dog Bite" is represented graphically in figure 4 The interpretation of the variables is as follows $E = 0$ or 1 depending on whether the explosion occurs, $P = 0$, 1, or 2, depending on whether Terrorist does not push the detonator at noon, pushes it with his right hand, or pushes it with his left hand (respectively), and $D = 0$ or 1 depending whether the dog bites his right hand or not, shortly before noon Note that the variable P is not binary it does not merely represent whether he pushes the detonator with his left hand or not, but represents whether he pushes the detonator at all, and if so with which hand The structural equations for this system are

---

[30] Note the similarity to the definition of *actual cause* in Pearl and Halpern

[31] The reader may verify that *WA* is not met in either "Dog Bite" or "Boulder"

$$\mathscr{DB} \quad D = 1,\ P = 1 + D,\ E = P \wedge 1$$

In words the dog does bite Terrorist's right hand shortly before noon, Terrorist will push the detonator with his right hand at noon unless the dog bites his right hand, in which case he will push the detonator with his left hand, and the bomb will explode just in case he pushes the detonator with his right or left hand at noon In fact, the values of these variables are $D = 1$, $P = 2$, $E = 1$ Suppose the dog had not bitten his hand Then we set $D = 0$, yielding the values $P = 1$ and $E = 1$ That is, if the dog had not bitten his right hand, Terrorist would have pushed the detonator with his right hand at noon and the bomb would still have exploded This system of equations yields the intuitively correct result that the explosion does not counterfactually depend upon the dog bite

The dog bite is not a cause of the explosion there is only one route from D to E, namely, $\langle D, P, E\rangle$ and that route is not causally active In disanalogy with "Backup," the failure of E to counterfactually depend upon D is not due to cancellation along different routes, but rather to what we might call a *failure of composition* While the function f_P from the variable D to P is nontrivial, and the function f_E from P to E is nontrivial, the composite function $f_E \circ f_P$ from D to E is trivial

L A Paul[32] argues that "Dog Bite" (or rather a case structurally identical to it) does not constitute a violation of the transitivity of causation She maintains that the effect of the dog bite is different from the cause of the explosion, one is a push qua thing done with left hand, the other a push qua push of detonator button I take this to be a proposal about how to use causal language to describe the pattern of dependence captured by $\mathscr{DB}$ Perhaps it is possible to regiment our causal language in such a way that failures of composition need not be described by triples of the form '*a* causes *b*', '*b* causes *c*', and '*a* does not cause *c*' But since the transitivity of causation is not a principle of which we have independent need (section v), and since there are counterexamples to the transitivity of

---

[32] "Aspect Causation," this JOURNAL, XCVII, 4 (April 2000) 235-56

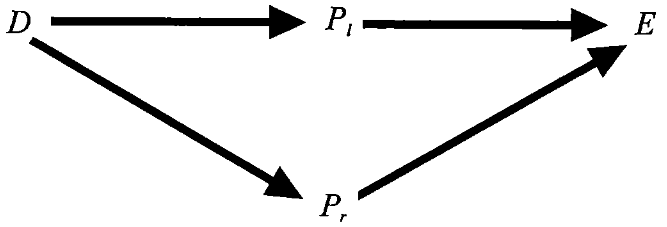


Figure 5

causation which do not rest on failures of composition (section VIII), the desire to preserve the transitivity of causation provides a poor motive for so regimenting our language

We have shown that there is no active path from D to E in the model $\langle\{D, P, E\}, \mathscr{DB}\rangle$ Perhaps there is another choice of $\langle\mathscr{V}, \mathscr{E}\rangle$ that is equally appropriate? Suppose, for example, that we represent the situation using the structure depicted in figure 5 Here $P_l = 0$ or 1 depending upon whether Terrorist pushes the button with his left hand at noon, and $P_r = 0$ or 1 depending upon whether Terrorist pushes the button with his right hand at noon The corresponding set of structural equations would be

$$\mathscr{DB}' \quad D = 1,\ P_l = D,\ P_r = \neg D,\ E = P_l \vee P_r$$

This structure gives us two routes from D to E Holding P_r fixed at its actual value of 0, E *does* counterfactually depend upon D This implies that there is an active route, $\langle D, P_l, E\rangle$, from D to E

What makes this an inappropriate model for "Dog Bite"?[33] One minimum criterion of adequacy for a model is that it entail only true counterfactuals As it turns out, the very counterfactual that reveals the active route $\langle D, P_l, E\rangle$ in this model is false, or at best indeterminate That ENF counterfactual is given that Terrorist did not push the detonator button with his right hand, if the dog had not bitten Terrorist's right hand, then (he would not have pushed the button with his left hand either, and) the bomb would not have exploded[34] Given that he did not push the button with his right hand, why should the dog bite make any difference to whether he pushes it with the left? He wanted the bomb to explode, would he not push the button with his left hand regardless of whether the dog bit his right? At any

---

[33] The following discussion has benefited especially from comments by Hausman and Sober

[34] Note that $\mathscr{DB}$ does not entail that this counterfactual is true or that it is false Substituting $P \neq 1$ does not yield a determinate solution for either P or E

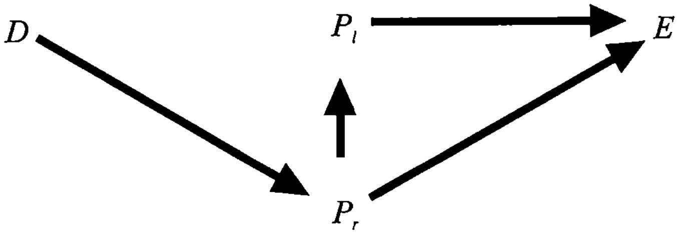


Figure 6

rate, it seems wrong to say that he would definitely *not* have pushed the button with his left hand

The problem with figure 5 and $\mathscr{DB}'$ is that they represent the dog bite as having two distinct effects it prevented Terrorist from pushing the button with his right hand, and caused him to push it with his left These effects are not genuinely distinct the dog bite caused Terrorist to push the button with his left hand only insofar as it prevented him from pushing it with his right This intuitive distinction is made precise using ENF counterfactuals Terrorist's pushing the button with his left hand ceases to depend counterfactually upon the dog bite when we specify whether or not he pushed the button with his right hand

We might alternately represent "Dog Bite" by $\mathscr{DB}''$ and figure 6

$$\mathscr{DB}'' \quad D = 1,\ P_r = \neg D,\ P_l = \neg P_r,\ E = P_l \vee P_r$$

This captures the idea that the dog bite causes Terrorist to push the button with his left hand *by* preventing him from pushing with his right

Figure 6 and $\mathscr{DB}''$ do not entail the troublesome ENF counterfactuals discussed above, but they are inappropriate for a different reason It was specified in the example that Terrorist had to press the detonator *at noon* in order for the bomb to explode Accordingly, the variables P_l and P_r must represent whether or not Terrorist presses the button with the relevant hand *at noon,* or else the equation for E will be false It is thus inappropriate to think of these variables as standing in a *causal* relationship Terrorist pushed the button at noon with his left hand rather than his right, but his failure to push the button with his right hand did not *cause* him to push the button with his left hand The two events (or rather, the one omission and the one event) fail to be distinct in the way required for the one to be a cause of the other Invoking the terminology of Lewis's new theory, pushing with the left hand and pushing with the right are alterations *of the same event* We capture this idea by representing the two as distinct values *of a single variable,* as we did in figure 4 and $\mathscr{DB}$

Figure 6 and $\mathscr{DB}''$ would be an appropriate representation for a slightly different case Suppose that there is a window of time during which Terrorist can push the button Suppose, moreover, that his initial instinct is to push the button with his right hand, and only after he finds himself unable to do so will he push the button with his left In this variant of the story, Terrorist's failure to push the button with his right hand does indeed cause him to push the button with his left hand a moment later In this version of the example, we still have the intuition that the dog bite did not cause the explosion, and I owe an account of this It turns out that this structure (or rather a crucial part of it) is isomorphic to that of "Boulder," which I discuss in the next section

Consider one final alternative representation Suppose we replace D with a three-valued variable the dog might do nothing, bite Terrorist's right hand, or maul Terrorist severely If the dog mauls Terrorist, he will not push the button (being totally incapacitated) and the bomb will not explode Now the route $\langle D, P, E\rangle$ in figure 4 will be active whether the bomb explodes will depend upon what the dog does But this is as it should be if we are willing to take seriously the possibility that the dog might have mauled Terrorist severely, we might well judge that the dog's (merely) biting his right hand (rather than mauling him) is a cause of the explosion [35] Assume, however, that this is not a possibility that we wish to take seriously (did *you* consider this possibility before I mentioned it?) then this new model is inappropriate for pragmatic reasons This mirrors Lewis's claim that in order for the event *c* to influence *e*, the alterations of *c* that yield different alterations of *e* must be "not-too-distant "

Of course, I cannot rule out every rival model of "Dog Bite " I maintain, however, that of those which spring to mind, only the model captured by figure 4 and $\mathscr{DB}$ is appropriate In this model, there is no active route from the dog bite to the explosion This captures our intuitive judgment that the dog bite is not a cause of the explosion In the course of discussing this example, we have illustrated three different ways in which a model might fail to be appropriate it might license counterfactuals that are not true, it might posit causal connections between events that are not distinct, or it

---

[35] This is the sort of case where the contrastive approach that I have developed elsewhere (see the references in footnote 26 above) is helpful

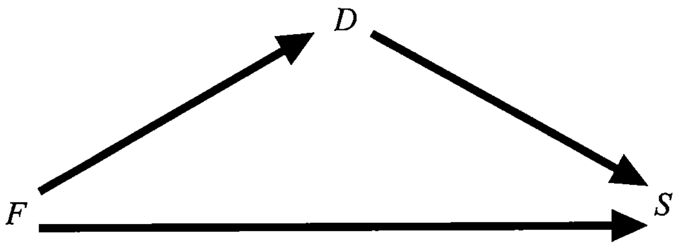


Figure 7

might encompass possibilities that are too distant from actuality to take seriously

VIII "BOULDER" REVISITED

Let us set up a model for "Boulder" (figure 7) The variable $F = 0$ or 1 depending upon whether the boulder falls, $D = 0$ or 1 depending upon whether Hiker ducks, and $S = 0$ or 1 depending upon whether Hiker survives The structural equations are

$$\mathcal{BO}\cdot\ F = 1,\ D = F,\ S = \neg F \vee D$$

(To see the isomorphism with $\mathcal{DB}''$, identify P_r with $\neg F$, P_l with D, and E with S)

In this model, there are no active routes from F to S It is relatively easy to see that the direct route $\langle F, S\rangle$ is inactive Holding D fixed at its actual value of 1, there is no counterfactual dependence of S on F holding fixed that Hiker ducked, he would have survived if the boulder had not fallen The more important issue, however, is whether the indirect route $\langle F, D, S\rangle$ is active It is along this route that we have a chain of counterfactual dependence, and hence causation on Lewis's account There are no intermediates between F and S that do not lie along this route, and hence nothing to hold fixed in evaluating whether this route is active Since S does not depend on F, the route is inactive Intuitively, the falling boulder does not save Hiker's life because without it, Hiker's life would not have been endangered in the first place This is just what is indicated by the absence of an active route from F to S there is no scenario in which the boulder does not fall and Hiker does not survive

Note that while the causal diagram for "Boulder" is isomorphic to that in "Backup" (figure 3), the structural equations are not isomorphic Thus "Backup," "Dog Bite," and "Boulder" all exhibit different causal structures counterfactual dependence fails for different reasons in each case By contrast, Lewis[36] claims that all of the (putative) counterexamples to transitivity have a common structure This is not surprising it is only by considering ENF counterfactuals that the difference in structure is revealed If we consider only counterfactuals with single antecedents, as Lewis does, then all three cases have the same structure there is a chain of counterfactual dependence, but no dependence of the last event on the first

There is a sense in which the inactivity of the route $\langle F, D, S \rangle$ rests on a loophole in the definition of an active route The guiding idea behind that definition was to isolate the influence of one variable on another along a given route by factoring out any counteracting influences along other routes This was accomplished by holding fixed the values of intermediate variables along those other routes This technique will not work, however, when one of the other routes in question is direct a direct route has no intermediate variables, so its role cannot be factored out I maintain that this is no weakness in the definition, rather, it is the nature of direct routes that they are too intimately bound up with other routes to be separated from them Any effect the falling boulder may have made on Hiker's survival by causing him to duck cannot be conceptually severed from the risk that the falling boulder posed to him in the first place Thus "Boulder" is analogous to "Backup" in that S fails to depend counterfactually upon F because of cancellation along two different routes, but "Boulder" is unlike "Backup" in that an active route cannot be isolated

Is it possible to interpolate a variable along the route $\langle F, S \rangle$? If so, then it may be possible to hold *that* variable fixed, and thus isolate the route $\langle F, D, S \rangle$ and show it to be active This can be done, but it turns out to be much more difficult than it might first appear The trick is to find a variable that does not also lie along the route $\langle F, D, S \rangle$ The possibility of doing this rests upon the contingent fact that it takes Hiker a finite amount of time to react to the boulder and get into a safe position There will be a point on the boulder's trajectory—let us say one meter from Hiker's head—such that by the time the boulder reaches that point, it is too late for Hiker to duck if he has not done so already Let $B = 1$ represent the boulder's presence at this point, and $B = 0$ its absence Let us add the clarifying remark that $D = 0$ or 1 according to whether Hiker ducks *at the appropriate time*—sufficiently early to avoid being hit by the boulder The value of D tells us nothing about whether Hiker attempts (too late) to duck in response

---

[36] "Causation as Influence," pp 194-95

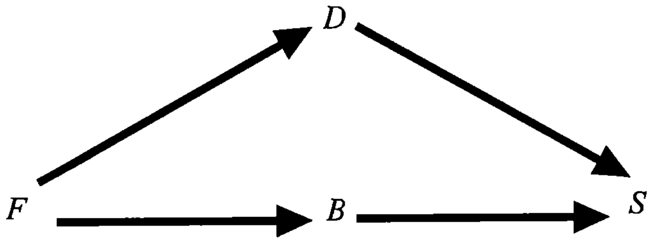


Figure 8

to the boulder when it is only one meter from his head Note how we have carefully chosen B so that neither B nor D affect each other Here is the new set of structural equations

$$\mathscr{BO}' \quad F = 1,\ D = F,\ B = F,\ S = \neg B \vee D$$

The corresponding causal graph is shown in figure 8 In fact, B took the value 1 Holding this fixed, we must determine whether Hiker's survival depends counterfactually upon the boulder's fall Setting $F = 0$ and $B = 1$, we get $D = 0$, and thus $S = 0$ Relative to the system of structural equations $\mathscr{BO}'$, the route $\langle F, D, S\rangle$ is indeed active

Must we then conclude that the boulder's fall did save Hiker's life after all? Perhaps this would not be so bad On my account, unlike Lewis's, it is at least easy to explain why our pretheoretic intuition yields the "wrong" answer in this case The interpolated variable B is not easy to find, and the ENF counterfactual that reveals the active causal route from F to S is not at all intuitive In words, the relevant piece of counterfactual reasoning would go as follows suppose that the boulder had been present at a point one meter from Hiker's head and flying toward him, and suppose moreover that it had never fallen in the first place Since it never fell, Hiker would not have seen it coming and would not have ducked, since it would have been there, one meter from his exposed head, it would have hit him and he would not have survived This counterfactual reasoning is correct, but bizarre If the boulder never fell, how did it get to be there, one meter from Hiker's head? We are to imagine, presumably, that the boulder was mysteriously and instantaneously transported to a position immediately in front of Hiker's head This is the sort of counterfactual reasoning that only trained philosophers engage in, unaided intuition is not to be faulted for failing to "see" the relevant ENF counterfactual

I believe, however, that it is possible to give a less apologetic response [37] Figure 8 and $\mathscr{BO}'$, while more complete than figure 7 and $\mathscr{BO}$, do not constitute an appropriate representation of "Boulder " As we have noted, our causal judgments depend, in part, upon which unactualized possibilities we are willing to take seriously, and which we consider too remote The variables we choose to include in a causal model should reflect these concerns We included the variables F and D in our original model of "Boulder " This choice reflects our willingness to take seriously the possibility that the boulder does not fall, and the possibility that Hiker does not duck Moreover, it reflects our willingness to take seriously the possibility Hiker does not duck *even though the boulder falls* That is, it reflects our willingness to view Hiker's life as being at risk Why are we willing take this possibility seriously, even though it is stipulated that Hiker is determined to duck if the boulder falls? Plausibly, it is because we recognize the nomic connection between the boulder's fall and Hiker's duck to be highly contingent upon unspecified details of the case Hiker might not have ducked if he had been looking in the wrong direction, if his reactions were slowed by tired muscles, if he were less prone to reacting coolly in times of crisis, and so on It is not necessary, however, that we have any one of these explanations firmly in mind in order to take seriously the possibility that Hiker does not duck when the boulder falls

When we exclude the variable B from our model, it is not because we are unwilling to take seriously the possibility that the boulder was not present at that point (one meter from Hiker's head) We take that possibility seriously when we entertain the possibility that the boulder does not fall in the first place Rather, we are not willing to take seriously the possibility that the boulder (or *a* boulder of similar size and shape) comes to be in that position *even though the boulder does not fall in the first place* This possibility is just too far-fetched (Did *you* consider this possibility before I mentioned it?) Perhaps one could tell a story that would lead us to take this possibility seriously—perhaps Hiker has inadvertently walked in front of a boulder launcher that is carefully camouflaged against the hillside [38] But in just such a case, we should take the original causal claim seriously by causing Hiker to duck in plenty of time, the fall of the boulder down the hillside does indeed save Hiker's life

---

[37] The following paragraphs owe a great deal to discussion with Pearl

[38] To get the causal structure to come out right, more details would be needed In particular, it would have to be the case that on this occasion, the launching of a boulder is triggered by the original boulder falling down the hillside

IX CONCLUSION

If one subscribes to the thesis that causation is transitive, there are benefits one can explain how there can be causation without counterfactual dependence in cases of preemption such as “Backup ” But there are costs as well the defender of transitivity is committed to unintuitive causal claims in cases like “Dog Bite” and “Boulder ” The benefits are not worth the costs they can be had cost-free The use of ENF counterfactuals allows us to accommodate causation without counterfactual dependence in cases of preemption, without committing us to the counterintuitive consequences of transitivity The use of ENF counterfactuals is independently motivated by the techniques employed in causal modeling

In addition to reproducing our intuitions in standard test cases, the approach recommended here has a number of further advantages It reproduces our intuitive causal judgments in a very natural way the counterfactuals appealed to are psychologically natural counterfactuals rather than philosophically technical ones Moreover, it reproduces our causal judgments without introducing events or variables beyond those explicitly presented in the various scenarios (such as intermediate stages in the trajectory of the bullet in “Backup” or the boulder in “Boulder”) In short, it reproduces our intuitive judgments using only those resources that are available to unaided intuition The approach recommended here allows for an elegant representation of patterns of counterfactual dependence We can explicitly represent the structural differences between superficially similar cases These representations also allow us to make explicit some of our underlying assumptions about a particular case (such as that certain possibilities are too far-fetched to be taken seriously)

Perhaps there is still room for the defender of transitivity to maneuver But when there is an alternative account with such striking advantages, why bother?

CHRISTOPHER HITCHCOCK

California Institute of Technology

PARADOXES OF FAIR DIVISION*

Paradoxes, if they do not define a field, render its problems intriguing and often perplexing, especially insofar as the paradoxes remain unresolved Voting theory, for example, has been greatly stimulated by the Condorcet paradox, which is the discovery by the Marquis de Condorcet[1] that there may be no alternative that is preferred by a majority to every other alternative, producing so-called cyclical majorities Its modern extension and generalization is Arrow's[2] theorem, which says, roughly speaking, that a certain set of reasonable conditions for aggregating individuals' preferences into some social choice are inconsistent

In the last fifty years, hundreds of books and thousands of articles have been written about these and related social-choice paradoxes and theorems, as well as their ramifications for voting and democracy Hannu Nurmi[3] provides a good survey and classification of voting paradoxes and also offers advice on "how to deal with them "

There is also an enormous literature on fairness, justice, and equality, and numerous suggestions on how to rectify the absence of these properties or attenuate their erosion But paradoxes do not frame the study of fairness in the same way they have inspired social-choice theory

To be sure, the notion that justice and order may be incompatible, or that maximin justice in John Rawls's[4] sense undercuts the motivation of individuals to strive to do their best, underscores the possible trade-offs in making societies more just or egalitarian For example, an egalitarian society may require strictures on free choice to ward off anarchy, rewarding the worst-off members of a society may deaden competition among the most able if their added value is siphoned off to others

Obstacles like these which stand in the way of creating a just society are hardly surprising They are not paradoxes in the strong sense of

---

[*] Brams acknowledges the support of the C V Starr Center for Applied Economics at New York University Research by Edelman was done while he was in the School of Mathematics, University of Minnesota We thank Dorothea Herreiner for valuable comments

[1] *Essai sur l'application de l'analyse a la probabilité dés decisions rendues a la pluralité des voix* (Paris De L'Imprimerie royale,1785)

[2] Kenneth J Arrow, *Social Choice and Individual Values* (New Haven Yale, 1951, 2nd ed , 1963)

[3] *Voting Paradoxes and How to Deal with Them* (Berlin Springer, 1999)

[4] *A Theory of Justice* (Cambridge Harvard, 1971)

0022-362X/01/9806/300-14

constituting a logical contradiction between equally valid principles Here we use paradox in a weaker sense—as a conflict among fairness conditions that one might expect to be compatible Because we are surprised to discover this conflict, it is "nonobvious," as one of us labeled a collection of paradoxes he assembled about politics [5]

The fair-division paradoxes we present here all concern how to divide up a set of indivisible items among two or more players In some paradoxes, we assume the players can do no more than rank the items from best to worst, in others, we assume they can, in addition, indicate preferences over subsets, or packages, of items While our framework is generally an ordinalist one, we do admit one cardinalization of ranks, based on the Borda count used in voting, to facilitate certain comparisons, particularly those involving allocations with different numbers of items

The main criteria we invoke are *efficiency* (there is no other division better for everybody, or better for some players and not worse for the others) and *envy-freeness* (each player likes her allocation at least as much as those which the other players receive, so she does not envy anybody else) But because efficiency, by itself, is not a criterion of fairness (an efficient allocation could be one in which one player gets everything and the others nothing), we also consider other criteria of fairness besides envy-freeness, including two different measures of how a worst-off player fares (maximin and Borda maximin), which we contrast with a utilitarian notion of overall welfare (Borda total score) What we rule out, besides the splitting of items, is the possibility of randomizing among different allocations, which is another way that has been proposed for "smoothing out" inequalities caused by the indivisibility of items [6]

---

[5] Brams, *Paradoxes in Politics An Introduction to the Nonobvious in Political Science* (New York Free Press, 1976), see also Fishburn, "Paradoxes of Voting," *American Political Science Review*, LXVIII, 2 (June 1974) 537-46, and Fishburn and Brams, "Paradoxes of Preferential Voting," *Mathematics Magazine*, LVI, 4 (September 1983) 207-14

[6] Fair-division procedures that allow for the splitting of (divisible) goods or the sharing of (indivisible) goods—possibly based on a randomization process that determines time shares—are discussed in, among other places, John Broome, *Weighing Goods Equality, Uncertainty and Time* (Cambridge Blackwell, 1991), H Peyton Young, *Equity in Theory and Practice* (Princeton University Press, 1994), Hervé Moulin, *Cooperative Economics A Game-Theoretic Introduction* (Princeton University Press, 1995), Brams and Alan D Taylor, *Fair Division From Cake-Cutting to Dispute Resolution* (New York Cambridge, 1996), Jack Robertson and William Webb, *Cake-Cutting Algorithms Be Fair If You Can* (Natick, MA A K Peters, 1998), and Brams and Taylor, *The Win-Win Solution Guaranteeing Fair Shares to Everybody* (New York Norton, 1999)

Our paradoxes demonstrate the opportunities as well as the limitations of fair division Thus, for example, while the only division of items in which one player never envies the allocation of another may be nonexistent or inefficient, we note that there is always an efficient and envy-free division for two players—even when they rank all items the same—as long as they do not rank all subsets of items the same We also show that fair division may entail an unequal division of the items

We divide the paradoxes into three categories

(1) The conflict between efficiency and envy-freeness (paradoxes 1 and 2)
(2) The failure of a unique efficient and envy-free division to satisfy other criteria (paradoxes 3 and 4)
(3) The desirability, on occasion, of dividing items unequally (paradoxes 5, 6, 7, 8)

While the paradoxes highlight difficulties in creating "fair shares" for everybody, they by no means render the task impossible Rather, they show how dependent fair division is on the fairness criteria one deems important and the trade-offs one considers acceptable Put another way, achieving fairness requires some consensus on the ground rules (that is, criteria) and some delicacy in applying them (to facilitate trade-offs when the criteria conflict)

We mention three technical points before we proceed to specific examples First, we assume that players cannot compensate each other with side payments—the division is only of the indivisible items Second, all players have positive values for every item Third, a player *prefers* one set *S* of items to a different set *T* if (i) *S* has as many items as *T* and (ii) for every item *t* in *T* and not in *S*, there is a distinct item *s* in *S* and not *T* that the player prefers to *t* For example, if a player ranks items 1 through 4 in order of decreasing preference 1234, we assume that she prefers

- the set {1,2} to {2,3}, because {1} is preferred to {3}, and
- the set {1,3} to {2,4}, because {1} is preferred to {2} and {3} is preferred to {4}

whereas the comparison between sets {1,4} and {2,3} could go either way

I EFFICIENCY AND ENVY-FREENESS THEY MAY BE INCOMPATIBLE

Paradox 1 a unique envy-free division may be inefficient Suppose there is a set of three players, {*A*, *B*, *C*}, that must divide a set of six indivisible items, {1, 2, 3, 4, 5, 6} Assume the players strictly rank the items from best to worst as follows

Example I
A 1 2 3 4 5 6
B 4 3 2 1 5 6
C 5 1 2 6 3 4

The *unique* envy-free allocation to (*A*, *B*, *C*) is ({1,3}, {2,4}, {5,6}), or for simplicity (13, 24, 56), whereby *A* and *B* get their best and third-best items, and *C* gets her best and fourth-best items Clearly, *A* prefers her allocation to that of *B* (which are *A*'s second-best and fourth-best items) and that of *C* (which are *A*'s two worst items) Likewise, *B* and *C* prefer their allocations to those of the other two players Consequently, the division (13, 24, 56) is envy-free all players prefer their allocations to those of the other two players, so no player is envious of any other

Compare this division with (12, 34, 56), whereby *A* and *B* receive their two best items, and *C* receives, as before, her best and fourth-best items This division *Pareto-dominates* (13, 24, 56), because two of the three players (*A* and *B*) prefer the former allocation, whereas both allocations give player *C* the same two items (56)

It is easy to see that (12, 34, 56) is Pareto-optimal, or efficient no player can do better with some other division without some other player or players doing worse This is apparent from the fact that the only way *A* or *B*, who get their two best items, can do better is to receive an additional item from one of the two other players—assuming all items have some positive value for the players—but this will necessarily hurt the player who then receives fewer than her present two items Whereas *C* can do better without receiving a third item if she receives item 1 or 2 in place of item 6, this substitution would necessarily hurt *A*, which will do worse if she receives item 6 for item 1 or 2

The problem with efficient allocation (12, 34, 56) is that it is not *assuredly* envy-free In particular, *C* will envy *A*'s allocation of 12 (second-best and third-best items for *C*) if she prefers these two items to her present allocation of 56 (best and fourth-best items for *C*) In the absence of information about *C*'s preferences for subsets of items, therefore, we cannot say that efficient allocation (12, 34, 56) is envy-free [7]

---

[7] Henceforth we shall mean by 'envy-free' a division such that, no matter how the players value subsets of items consistent with their rankings, no player prefers any other player's allocation to her own If a division is not envy-free, we call it *envy-possible* if a player's allocation *may* make her envious of another player, depending on how she values subsets of items, as illustrated by division (12, 34, 56) in the text It is *envy-ensuring* if it causes envy, independent of how the players value subsets

But the real bite of paradox 1 stems from the fact that not only is inefficient division (13, 24, 56) envy-free, but it is uniquely so—there is no other division, including an efficient one, that guarantees envy-freeness To show this in example I, note first that an envy-free division must give each player her best item, if not, then a player might prefer a division, like envy-free division (13, 24, 56) or efficient division (12, 34, 56), that does give each player her best item, rendering the division that does not envy-possible or envy-ensuring Second, even if each player receives her best item, this allocation cannot be the only item she receives, because then the player might envy any player that receives two or more items, *whatever* these items are

By this reasoning, then, the only possible envy-free divisions in example I are those in which each player receives two items, including her top choice It is easy to check that no efficient division is envy-free[8] Similarly, one can check that no inefficient division, except (13, 24, 56) that gives each player two items—including her best—is envy-free, making this division uniquely envy-free

Paradox 2 there may be no envy-free division, even when all players have different preference rankings While it is bad enough when the only envy-free division is inefficient (paradox 1), it seems even worse when there is no envy-free division This is trivial to show when players rank items the same For example, if two players both prefer item 1 to item 2, then the player that gets item 2 will envy the player that gets item 1

In the following example, each of three players has a different ranking of three items

Example II
A 1 2 3
B 1 3 2
C 2 1 3

There are three divisions, in which each player gets exactly one item, which are efficient—(1, 3, 2), (2, 1, 3), and (3, 1, 2)—in each of which at least one player gets her best item It is evident that none is

---

of items In effect, a division that is envy-possible has the potential to cause envy By comparison, an envy-ensuring division always causes envy, and an envy-free division never causes envy

[8] We previously showed that division (12, 34, 56) is not envy-free As another example, consider efficient division (16, 34, 25) Whereas neither *B* nor *C* envies each other or *A*, *A* might prefer either *B*'s 34 or *C*'s 25 allocations, making this division envy-possible

envy-free, because the player that gets item 1 in each (*A* or *B*) will be envied by at least one of the other two players. For instance, in the case of the division (2, 1, 3), both *A* and *C* will envy *B*

Can an inefficient division be envy-free, as was the case in example I? It is not hard to see that this situation cannot occur in example II for the reason given above the player that gets item 1 will be envied But in the case of an inefficient division, "trading up to efficiency" reduces the *amount* of envy For example, consider inefficient division (2, 3, 1), in which each player receives her second-best choice Because *A* envies *C*, *B* envies *C*, and *C* envies *A*, a trade of items 1 and 2 between *A* and *C* is possible It yields efficient division (1, 3, 2), in which only *B* envies *A*

Besides (2, 3, 1), the other two inefficient divisions—(1, 2, 3) and (3, 2, 1)—also allow for trading up to efficiency In the first, a trade of items 2 and 3 between *B* and *C* yields efficient division (1, 3, 2), in the second, a trade of items 1 and 2 between *B* and *C* yields efficient division (3, 1, 2) Three-way trades are also possible For instance, starting from inefficient division (3, 2, 1), a three-way trade, whereby *A* sends item 3 to *B*, *B* sends item 2 to *C*, and *C* sends item 1 to *A*, yields efficient division (1, 3, 2)

Trading up to efficiency is also possible in example I by exchanging items 2 and 3, *A* and *B* can turn inefficient division (13, 24, 56) into efficient division (12, 34, 56) As in example II, however, no efficient division is envy-free in example I The difference between examples I and II is that example II does not admit even an inefficient envy-free division

II UNIQUE EFFICIENT AND ENVY-FREE DIVISIONS THEIR INCOMPATIBILITY WITH OTHER CRITIERIA

Paradox 3 a unique efficient and envy-free division may lose in voting to an efficient and envy-possible division So far we have shown that efficiency and envy-freeness may part company either by there being no envy-free division that is also efficient (example I), or no envy-free division at all (example II) But when these properties coincide, and there is both an efficient and an envy-free division, it may not be the choice of a majority of players, as illustrated by the following example

Example III
A 1 2 3 4 5 6
B 5 6 2 1 4 3
C 3 6 5 4 1 2

There are three efficient divisions in which (*A*, *B*, *C*) each get two items (12, 56, 34), (12, 45, 36), and (14, 25, 36) But only the third division, (14, 25, 36), is envy-free Whereas *C* might prefer *B*'s 56 allocation in the first division, and *B* might prefer *A*'s 12 allocation in the second division, no player prefers another player's allocation in (14, 25, 36)

But observe that both *A* and *B* prefer the first division, (12, 56, 34), to the envy-free third division, (14, 25, 36), because they get their top two items in the first division, only *C* gets her top two items in (14, 25, 36) Hence, the first division would defeat the envy-free third division, (14, 25, 36), by simple majority rule

The situation is not so clear-cut when we compare the second division, (12, 45, 36), with the envy-free (14, 25, 36) In fact, there would be a tie vote *C* would be indifferent, because she gets her top two items, 36, in each division, *A* would prefer the second division (top two items versus best and fourth-best items), and *B* would prefer the envy-free division, (14, 25, 36) (best and third-best items versus best and fifth-best items)

Thus, if there were a vote, the unique envy-free division, (14, 25, 36), would lose to the envy-possible division, (12, 56, 34), and it would tie with the other envy-possible division, (12, 45, 36) If there were "approval voting,"[9] and *A*, *B*, and *C* voted only for the divisions that give each player her two best items, then the envy-free division, (14, 25, 36), would get 1 vote, compared to 2 votes each for both of the envy-possible divisions, (12, 56, 34) and (12, 45, 36) In sum, players will choose an envy-possible over the unique envy-free division, (14, 25, 36), in either pairwise comparisons or approval voting

Paradox 4 neither the Rawlsian maximin criterion nor the Borda total-score criterion may choose a unique efficient and envy-free division Besides using voting to select an efficient division, consider the following Rawlsian maximin criterion to distinguish among efficient divisions choose the division that maximizes the minimum rank of items that players receive, making a worst-off player as well off as possible To illustrate in example III, envy-possible division (12, 45, 36) gives a fifth-best item to *B*, whereas each of the two other efficient divisions gives a player, at worst, a fourth-best item Between the latter two divisions, the envy-possible division, (12, 56, 34) is, arguably, better than the envy-free division, (14, 25, 36), because it gives the other two players—those which do not get a fourth-best item—their two best items, whereas envy-free division (14, 25, 36) does not give *B* her two best items [10]

---

[9] Under approval voting, voters can vote for as many alternatives as they like, each alternative approved of receives one vote, and the alternative with the most votes wins See Brams and Fishburn, *Approval Voting* (Boston Birkhauser, 1983)

A modified Borda count would also give the nod to the envy-possible division, (12, 56, 34), compared not only with the envy-free division, (14, 25, 36), but also with the other envy-possible division, (12, 45, 36) Awarding 6 points for obtaining a best item, 5 points for obtaining a second-best item, , 1 point for obtaining a worst item in example III, the latter two divisions give the players a total of 30 points, whereas envy-possible division (12, 56, 34) gives the players a total of 31 points,[11] which we call their *Borda total scores* and use as a measure of the *overall* utility or welfare of the players Hence an envy-possible division beats the unique envy-free division, based on both the maximin criterion and the Borda total-score criterion (We shall later apply the Borda count to individual players, asking what division maximizes the minimum Borda score that any player receives)

III THE DESIRABILITY OF UNEQUAL DIVISIONS (SOMETIMES)

Paradox 5 an unequal division of items may be preferred by all players to an equal division In section II, we showed that neither (i) pairwise comparison voting or approval voting (paradox 3), nor (ii) the maximin criterion or the Borda total-score criterion (paradox 4), always selects a unique efficient and envy-free division In the following example, there is also a unique efficient and envy-free division—in which all players receive the same number of items (henceforth called an *equal division*)—but there may be grounds for choosing an efficient but unequal envy-possible division

Example IV
A 1 2 3 4
B 2 3 4 1

---

[10] This might be considered a second-order application of the maximin criterion if, for two divisions, players rank the worst item any player receives the same, consider the player that receives a next-worst item in each, and choose the division in which this item is ranked higher This is an example of a *lexicographic decision rule*, whereby alternatives are ordered on the basis of a most important criterion, if that is not determinative, a next-most important criterion is invoked, and so on, to narrow down the set of feasible alternatives

[11] The standard scoring rules for the Borda count in this 6-item example would give 5 points to a best item, 4 points to a second-best item, , 0 points to a worst item We depart slightly from this standard scoring rule to ensure that each player obtains some positive value for all items, including her worst choice, as assumed earlier

It is not difficult to show that (13, 24) is the only efficient and envy-free division. Two other equal divisions, (12, 34) and (14, 23), while better for one player and worse for the other, are envy-possible.

The above three equal divisions all give Borda total scores of 12 to their players. If we eliminate the envy-possible division, (14, 23), on the grounds that it fails the maximin criterion by giving *A* her worst item (item 4), then the comparison reduces to that between envy-free division, (13, 24), and envy-possible division, (12, 34).

Curiously, it is possible that *both A* and *B* prefer the unequal envy-possible division, (134, 2), to the equal envy-possible division, (12, 34).[12] Thus, unequal divisions might actually be better for all players than equal divisions.

Ruling out equal division (12, 34) in such a situation, let us compare (134, 2) with the envy-free (equal) division (13, 24). Clearly, (134, 2) is better than (13, 24) for *A*, but it is worse for *B*.

This leaves open the question of which of these two divisions, involving an equal and an unequal division of the items, comes closer to giving the two players "fair shares." As the next paradox shows, an unequal division may actually be more egalitarian—as measured by Borda scores for individual players—than an equal division.

Paradox 6: an unequal division of items may (i) maximize the minimum Borda scores of players (Borda maximin) and (ii) maximize the sum of Borda scores (Borda maxsum). In paradox 5, we showed that an unequal but envy-possible division of items may compare favorably with an equal and envy-free division. To make this kind of comparison more precise, consider the following example.

Example V
A 1 2 3 4 5 6 7 8 9
B 3 1 2 4 5 6 7 8 9
C 4 1 2 3 6 5 7 8 9

There are exactly two unequal divisions, (12, 357, 4689) and (12, 3589, 467), that maximize the minimum Borda scores of players, which are [17, 17, 17] for both divisions.[13] On the other hand, there are two equal divisions, (129, 357, 468) and (129, 358, 467), that maximize the minimum Borda scores of players, which are [18, 17, 16] for the first division and [18, 16, 17] for the second division. These are all the divisions whose Borda total scores are 51, which, it can be shown, is the maximal sum, or *Borda maxsum*, among all possible divisions (equal or unequal)

---

[12] This is true if *A* prefers 34 to 2, and *B* prefers 2 to 34.

[13] Henceforth we shall indicate the Borda scores of players [in brackets] to distinguish them from item allocations (in parentheses).

Notice that the worst-off player in the two unequal divisions garners 17 points (so does the best-off player, because the Borda scores of all players are the same), whereas the worst-off player in the two equal divisions receives fewer points (16) By the maximin criterion, but now based on Borda scores, the unequal divisions are more egalitarian We call this the *Borda maximin* criterion, which is especially useful in comparing equal and unequal divisions[14]

None of the four equal or unequal divisions is envy-free—all are envy-possible or envy-ensuring Likewise, all four divisions are "efficient-possible" in the sense that there may be a more efficient division, but this is not guaranteed Take, for example, the unequal division (12, 357, 4689) *B* or *C* might prefer *A*'s 12 allocation, just as *A* might prefer *B*'s or *C*'s allocation, so a trade could make two, or even all three, players better off Unlike our previous examples, in which divisions called "efficient" were all "efficient-ensuring" (that is, there were no trades that could improve the lot of all traders, however players valued subsets of items), this is not the case in example V

The Borda maximin criterion seems a reasonable one to distinguish among all efficient-possible and envy-possible divisions In example V, it is not only unequal divisions that do best on this criterion, but these divisions are also Borda maxsum, making them both fair and utility-maximizing (according to the Borda cardinalization of utility—more on its limitations, illustrated in example VII, later)

Paradox 7 an unequal division of items may be Borda maxsum but not Borda maximin There was no conflict between Borda maxsum and Borda maximin in example V—two unequal divisions satisfied both these properties But as the next example illustrates, this need not be the case

Example VI
A 1 2 3
B 1 3 2

---

[14] To see why, consider the two unequal divisions in example IV in which no player receives a fourth-best item (1, 234) and (123, 4) To call these divisions maximin—like equal division (13, 24), in which no player receives a fourth-best item as well—seems highly questionable, because the player receiving her top three items in these two divisions can hardly be considered worse off (because she receives a third-best item) than the player receiving only her top item Indeed, the Borda scores of the players in the two unequal divisions, [4, 9] and [9, 2], reveal how inegalitarian these divisions are, particularly when compared with equal Borda maximin division (13, 24) with Borda scores of [6, 6]

There are two unequal maxsum divisions, (12, 3) and (2, 13), whose Borda scores are, respectively, [5, 2] and [2, 5] Each gives a Borda total score of 7, and a minimum Borda score of 2 for a player

By contrast, there are two Borda maximin divisions, (1, 23) and (23, 1), both of which give Borda scores of [3, 3] While they give the players a lower Borda total score (6) than the Borda maxsum divisions, they give the players a higher minimum score of 3

Presumably, the egalitarian would choose one of the two Borda maximin divisions, whereas the utilitarian would choose one of the two Borda maxsum divisions Because there are an odd number of items to divide in example VI, all the divisions between *A* and *B* are necessarily unequal But both Borda maximin divisions and Borda maxsum divisions can be the either equal (unlike example VI, the Borda maximin division in example IV is equal, as discussed in footnote 14) or unequal (two of the Borda maxsum divisions in example V are equal and two are unequal) It also turns out that Borda maxsum and Borda maximin scores can be arbitrarily far apart[15]

We believe that when Borda maximin and Borda total scores choose different divisions, Borda maximin generally gives the fairer division by guaranteeing that the Borda score of the worst-off player is as great as possible[16] As we shall show in our final paradox, however, a Borda maximin division may be quite implausible, depending on how players value subsets, or it may not be envy-free when, at the same time, there exists an envy-free division that is neither maximin nor Borda maximin

Paradox 8 if there are envy-free divisions, none may be maximin or Borda maximin In the following example, there are two players but an odd number of items, so no equal division of the items is possible

Example VII
A 1 2 3 4 5
B 1 2 3 4 5

---

[15] Brams, Edelman, and Fishburn, "Fair Division of Indivisible Items" (unpublished)

[16] To be sure, assuming that the differences in ranks are all equal, as Borda scoring does, is a simplification If cardinal utilities could be elicited that reflect the players' intensities of preference, then these utilities—instead of the rank scores—could be used to equalize, insofar as possible, players' satisfaction with a division of the items For fair-division bidding schemes that incorporate cardinal information, see Brams and Taylor, *Fair Division* and *The Win-Win Solution*, Brams and D Marc Kilgour, "Competitive Fair Division," *Journal of Political Economy*, CIX, 2 (April 2001) 418-43, and Claus-Jochen Haake, Matthias G Raith, and Francis E Su, "Bidding for Envy-Freeness A Procedural Approach to *n*-Player Fair-Division Problems," *Social Choice and Welfare* (forthcoming)

Because the players rank the items exactly the same, all divisions are efficient, making the choice of a fairest one appear difficult

Only six divisions, however, are what Brams and Fishburn[17] call *undominated splits*

(1, 2345), (12, 345), (13, 245), (14, 235), (15, 234), (145, 23)

These divisions are those in which, in the absence of information about preferences over subsets, either of the two allocations in each might be preferred by a player, making each undominated All these divisions, therefore, are envy-possible

The Borda maximin divisions are (13, 245) and (14, 235), which give Borda scores of, respectively, [8, 7] and [7, 8] to the players But neither division might be envy-free if, say, *both* players prefer allocation 13 to 245 in the first and allocation 14 to 235 in the second—that is, both prefer the "same side" of each division These preferences imply that both prefer allocation 12 to 345 in the second division, (12, 345), and allocation 145 to 23 in the sixth division, (145, 23), precluding these divisions, as well, from being envy-free

Thus, the preferences of *A* and *B* assumed above would eliminate four of the undominated splits from being envy-free, allowing the two remaining divisions to be so For example, *A* might prefer allocation 1 in the first division and allocation 15 in the fifth, whereas *B* might prefer the complements allocation 2345 in the first, and allocation 234 in the fifth

In none of our previous examples with envy-free divisions was such a division not Borda maximin But as we have just illustrated, there may be several envy-free divisions, none of which is Borda maximin This divergence points to the limitation of Borda maximin as a criterion for choosing divisions, because Borda scoring may not reflect the intensity of player preferences that can be better gleaned from player preferences over subsets

But we do not need to know player preferences over subsets to show that an envy-free division may not be Borda maximin

Example VIII
A 1 2 3 4 5 6 7 8 9
B 5 8 1 2 6 7 3 4 9
C 3 4 9 1 2 5 6 7 8

---

[17] "Fair Division of Indivisible Items between Two People with Identical Preferences Envy-freeness, Pareto-optimality, and Equity," *Social Choice and Welfare*, XVII, 2 (February 2000) 247-67

It is easy to see that division (127, 568, 349) is envy-free, but it gives a seventh-best item to *A* By contrast, division (126, 587, 349) gives sixth-best items to *A* and *B* and the same 349 allocation to *C*

Because no other divisions (equal or unequal) give players lowest-ranked items that are as high as sixth-best, division (126, 578, 349) is maximin But it is not envy-free *B* may envy *A*, because she may prefer allocation 126 to 578, making this division envy-possible

The Borda scores of the envy-free division are [20, 22, 24], whereas those of the maximin division are [21, 21, 24], so the maximin division is also Borda maximin Both the envy-free and maximin/Borda maximin divisions have total Borda scores of 66, which is also the Borda maxsum in example VIII

This example illustrates what we think is our most striking paradox Specificially, without any special assumptions about the preferences of the players for subsets of items, it shows the clash between envy-freeness and both maximin and Borda maximin Furthermore, because there are no unequal divisions in example VIII that satisfy any of our fairness criteria—or, for that matter, the Borda maxsum criterion—it highlights the difficulty of choosing a fairest allocation, even in the equal-division case Should one help the worst-off, or avoid envy, when one cannot do both?[18]

---

[18] This question is more fully explored in Brams and Daniel King, "Efficient Fair Division Help the Worst Off or Avoid Envy?" (unpublished), in which it is shown not only that maximin and Borda maximin divisions may not be envy-free but also that *all* such divisions may actually ensure envy To illustrate this conflict, consider the following example

A 1 2 3 4 5 6
B 1 2 3 4 5 6
C 1 5 4 6 2 3

There are four Borda maximin divisions—(14, 23, 56), (23, 14, 56), (13, 24, 56), (24, 13, 56), each giving a minimum Borda score of 8 to a player—which are also maximin divisions (a worst-off player receives a fourth-best item) In addition, there are two maximin divisions which are not Borda maximin divisions—(12, 34, 56), (34, 12, 56)—which also give a worst-off player a fourth-best item All six divisions ensure envy in each, one player prefers another player's two items to her own This example demonstrates that maximin and Borda maximin divisions, rather than just preclude envy-freeness, may *guarantee* envy (that is, be envy-ensuring rather than just envy-possible, as in example VIII) Furthermore, unlike example II, in which the unique efficient maximin and Borda maximin division, (1, 3, 2), is also envy-ensuring, the present example involves each player's receiving two items, which one might think would be sufficient to allow a maximin or Borda maximin division to be envy-possible, if not envy-free This, however, is not the case, underscoring the seriousness of the conflict among our fairness criteria

IV CONCLUSIONS

The eight paradoxes pinpoint difficulties in dividing up indivisible items so that each player feels satisfied, in some sense, with her allocation The first two paradoxes show that efficient and envy-free divisions may be incompatible because the only envy-free division may be inefficient, or there may be no envy-free division at all

Both of these paradoxes require at least three players [19] When there are only two players, even when they rank items exactly the same, it turns out that efficient and envy-free divisions can always be found, except when the players have the same preferences over all subsets of items [20]

But the existence of even a unique efficient and envy-free division may not be chosen by the players for other reasons In particular, such a division will not necessarily be selected when players vote for the division or divisions that they prefer Also, a unique efficient and envy-free division will not necessarily be the division that maximizes the minimum rank of items that players receive, so the Rawlsian maximin criterion of making the worst-off player as well off as possible may not single it out

As a way of measuring the value of allocations to find those divisions which are most egalitarian, especially in comparing equal and unequal divisions, we used Borda scoring based on player rankings of the items We showed that a Borda maximin division may not be a Borda maxsum division, indicating the possible conflict between egalitarian and utilitarian outcomes

This difference may show up when there are as few as two players dividing up three items, making it impossible to divide the items equally between the players But even when this is possible, unequal divisions of items may be the only ones that satisfy the Borda maximin criterion While indicating a preference for this criterion over the Borda maxsum criterion when the two clash, we illustrated how Borda maximin divisions may fail badly in finding envy-free divisions Indeed, there may be no overlap between Borda maximin and envy-free divisions

Our purpose is not just to indicate the pitfalls of fair division by exhibiting paradoxes that can occur There are also opportunities, but these depend on the judicious application of selection criteria when not all criteria can be satisfied simultaneously

---

[19] Edelman and Fishburn, "Fair Division of Indivisible Items among People with Similar Preferences," *Mathematical Social Sciences*, XLI, 3 (May 2001) 327-47

[20] Brams and Fishburn, "Fair Division of Indivisible Items between Two People with Identical Preferences Envy-freeness, Pareto-Optimality, and Equity "

Several recent papers have suggested constructive procedures for finding the most plausible candidates for fair division of a set of indivisible items[21] We find this direction promising, because it is potentially applicable to ameliorating, if not solving, practical problems of fair division—ranging from the splitting of the marital property in a divorce to determining who gets what in an international dispute While some conflicts are ineradicable, as the paradoxes demonstrate, the trade-offs that best resolve these conflicts are by no means evident

STEVEN J BRAMS
New York University

PAUL H EDELMAN
Vanderbilt University

PETER C FISHBURN
AT&T Shannon Laboratory

---

[21] Brams and Fishburn, "Fair Division of Indivisible Items between Two People with Identical Preferences Envy-freeness, Pareto-Optimality, and Equity", Edelman and Fishburn, Brams, Edelman, and Fishburn, Brams and King, and Dorothea Herreiner and Clemens Puppe, "A Simple Procedure for Finding Equitable Allocations of Indivisible Goods," *Social Choice and Welfare* (forthcoming)

BOOK REVIEWS

Mind in a Physical World: An Essay on the Mind-body Problem and Mental Causation. JAEGWON KIM. Cambridge: Massachusetts Institute of Technology Press, 1998. vii + 146 p. Paper $15.00.*

During the previous century, philosophers of mind attempted to fashion a view of the relationship between the mental and the physical which is at once both physicalistic and nonreductive. At one time (and for some time), Jaegwon Kim thought that the concept of supervenience could be harnessed to spell out the relationship between the mental and the physical to this end. The idea was that, while the supervenience relation is tight enough to capture what is right about physicalism, it is also loose enough to be compatible with the irreducibility of the mental to the physical. The main message of *Mind in a Physical World* is that *nonreductive physicalism* (NRP) does not work. Kim argues that his and other versions of NRP inevitably run aground on the problem of mental causation. If he is right about this, then philosophers true to their physicalism will have to swallow reductionism. Those who find reductionism impossible to swallow will have to find a way of living without physicalism.

Kim sees the problem of mental causation as arising from the incompatibility of the following four plausible propositions:

(1) Many (or all) mental events cause physical events (alternatively: many mental properties are causally relevant to some physical properties).
(2) Mental events are distinct from physical events (the constitutive properties of mental events are not physical properties).
(3) The physical is nomologically and causally complete and closed.
(4) Physical events are not pervasively causally and/or nomologically overdetermined.

Each of these propositions taken singly is plausible, but together they are inconsistent. At least one of the four has to go. The question is: which?

Kim's discussion of NRP involves a number of ontological assumptions. One is a "sparse" view of properties on which properties are

---

* These are lectures that Kim originally gave as the Townsend Lectures at the University of California/Berkeley, in March 1996. This review is the better for having discussed these issues with Ned Block, Carsten Hansen, and Fritz Warfield and for having read drafts of their reviews.

0022-362X/01/9806/315-24

"mind-independent" features that "cut reality at its joints " On this view, not every predicate expresses a genuine property and it may be that not every property is expressed by a predicate of our language That properties cut nature at its joints means that they are involved in laws and causal relations in an appropriate way It follows, for example, that there is no guarantee that a disjunction of predicates that express properties will also express a property [1] If it turns out that a putative property is not involved in a law and/or causal relations in the right way, then it is not a genuine property According to Kim, an event is an instantiation of a property by an individual (or a relation by the appropriate number of individuals) at a time A mental event is the instantiation of a mental property For physicalism to have content, something must be said about the issue of characterizing the physical Kim's view seems to be that the microphysical properties of ideal physics are physical [2] He also counts as physical properties that are conjunctions and aggregates of microphysical properties and higher-level properties defined over lower-level physical properties (115) [3] If reductionism is true, then those mental predicates which express genuine properties express physical properties

The claim that physics is causally and nomologically closed and complete means that each physical event has a sufficient—or as sufficient as possible—causal/nomological explanation in terms of prior physical events The way I understand it, the causal/nomological completeness of physics is a contingent claim whose exact content depends on how 'physical event' is understood I think most physicists would agree that there is strong evidence in its favor when 'physical event' is understood as "event describable in the language of ideal micro-physics "[4]

NRP is a family of views differing in their ontological frameworks (views about properties, events, laws, and so on) and how they understand 'reduction' and 'physicalism' Following Kim, I understand the irreducibility of the mental to the physical as meaning that there exist mental events and properties that are distinct from any physical events and properties[5] Kim observes that a necessary condition for physicalism is that mental properties, events, and laws supervene on physical ones He considers various understandings of 'supervenience', but I think that any kind of physicalism worthy of the name requires a supervenience condition I shall call *Jackson physicalism* (JP) formulated as

---

[1] In some places, Kim suggests not only that genuine properties are always involved in causal relations, but the stronger view that (some of) their causal relations are essential to them

[2] I am not sure how one should characterize "ideal physics" except that it is safe to assume that the language of ideal physics contains no intentional or phenomenal predicates I have a bit more to say about this in "From Physics to Physicalism," in Carl Gillett and myself, eds , *Physicalism and Its Discontents* (New York Cambridge, 2001), pp 37-56

[3] Kim thinks that his characterization results in chemical and biological properties counting as physical, but it does not Being a water molecule is not an aggregate or conjunction of fundamental microphysical properties but a vast disjunction, since water molecules can occupy infinitely many quantum states

[4] This is Steven Weinberg's view in *Dreams of a Final Theory* (New York Random House, 1992)

> JP any minimal physical duplicate of the actual world is a duplicate *simpliciter*[6]

Some complications aside, JP entails that true mental propositions—for example, Jaegwon is thinking about sailing—are metaphysically entailed by true physical propositions (including propositions expressing physical laws) Kim thinks that supervenience relations are too weak to capture the idea that mental-property instantiations depend on physical-property instantiations, so he adds that the mental *depends* on the physical[7] This requirement threatens to make the supervenience condition superfluous and leaves dependence to be explicated Perhaps dependence can be captured by adding that the physical facts are the only natural class of facts which satisfies the JP scheme

The conflict among (1-4) is obvious If *P* occurs at *t*, the physical state *S* (which may consist of a fusion of many microphysical events) at *t-k* fully accounts for *P* If *M* supervenes on *S* and is a cause of *P*, then it appears that *P* is overdetermined by *S* and *M* If there is no such overdetermination, then either *M* does not cause *P* or is identical to some part of *S* If the latter, then it is a physical event and so

---

[5] Irreducibility is often taken also to have an epistemic component For example, the connections between physical and mental facts (properties, events) is enormously complex, nonexplanatory, epistemically inaccessible, and so on

[6] Frank Jackson—*From Metaphysics to Ethics* (New York Oxford, 1998)—gives this necessary condition on physicalism It captures the idea that all God has to do to create our world is to create its physical part, metaphysical connections take care of the rest That it is a necessary condition follows from the fact that, if it failed, then God would have more work to do after creating the physical facts and laws

[7] Two of the reasons that Kim gives for holding that supervenience is too weak to capture physicalism have to do with features of his formulation of supervenience One is that he formulates the doctrine as involving properties That leaves the possibility that there are laws and other facts that do not supervene on the physical Another is that he allows that the notion of necessity involved in supervenience is nomological necessity That allows for the possibility of fundamental bridge laws linking physical and mental properties and the existence of such laws would surely violate physicalism

(2) fails René Descartes's resolution of the problem of mental causation rejects the completeness of physics But this is not an option for physicalists since it involves there being nomological or causal relations that do not metaphysically supervene on the physical Epiphenomenalists reject (1) Given Kim's view about properties, that is tantamount to the view that mental predicates fail to express genuine properties, that is, to eliminativism Reductive physicalists reject (2) But there are persuasive and widely accepted arguments that mental predicates do not express physical properties Given Kim's view of events, this means that mental events are not physical events It looks like the only way out is to reject (4), but, as we shall see, Kim thinks there are reasons against that as well It is not for nothing that Arthur Schopenhauer calls the mind-body problem "the world knot " Kim's claim is that NRP is the latest failed attempt to untie it

In a nutshell, Kim's argument against NRP is this NRP is committed to (1-3) (1-3) imply the denial of (4) But (4) is more plausible than the conjunction of (1-3) So one of (1-3) is false

Kim thinks that escape from the problem of mental causation by rejecting (4) is not a reasonable option In *Mind in a Physical World*, he gives three reasons for this (44, 45) (A) it is implausible that every physical event that has a mental cause is causally overdetermined, (B) given the causal completeness and closure of physics mental causes are dispensable, (C) overdetermination may come into conflict with physical causal closure In addition to these arguments, Kim seems to suggest that (4) is a conceptual truth But so far as I can see, he provides no argument for this claim other than the preceding reasons and they, even if successful, fall short of establishing that

A typical case of causal overdetermination is the example of two assassins each independently shooting at the victim In this kind of example, the causes are wholly distinct events *independent* of each other Call this *independent overdetermination* The kind of case relevant to NRP is, in contrast, one in which the overdetermining causes are metaphysically connected to each other An example of this kind is a neurophysiological event N realizing a mental event M (say, a desire for a sip of beer) and where both cause E (say the subject's hand's moving) Call this kind *dependent overdetermination* Note that, in both cases, the causes by themselves do not determine the effect but, at most, determine for the effect given the laws and further circumstances Independent overdetermination is possible but rare If our evidence is that the prime minister has been shot, it will usually be more reasonable to posit a single assassin rather than two assassins acting independently The reason that we think the first kind of overdetermination is generally not plausible is that, when we posit two *independent* causes C1 and C2 for *E*, we raise the question of why both C1 and C2 and why the events that occurred led to the same effect. But in the case of causes that are metaphysically connected, we do not face these questions.[8] So Kim's remark that pervasive overdetermination of certain physical events by mental and physical events is implausible carries no force against NRP.

Kim's suggestion that mental causes are "dispensable" fares no better. The only way *M* could be dispensed with is by dispensing with *P* and *P* may not be dispensable with respect to *E*, that is, if *P* had failed to occur, *E* would have failed to occur. What Kim may have in mind is not that *M* is dispensable but that its being a *cause* of *E* is dispensable. Kim may have something like the following argument in mind. There are two possible worlds *w* and *w** in which (2) and (3) hold and which are exactly alike physically and mentally but differ only in that in *w* some mental events are overdetermining causes of certain physical events and in *w** mental events are epiphenomenal. Our evidence concerning causal relations would be exactly the same in *w* and *w**. Since *w* contains facts over and above *w** which are explanatorily idle with respect to the evidence, we should believe that the actual world is *w** rather than *w*. This argument seems to be appealing to the principle 'Don't multiply causes without necessity'. Whatever the merits of this principle in general, it is not clear that it applies to this case since the two causes are necessarily connected.

Kim's final argument is this:

> For consider a world in which the physical cause does not occur and which in other respects is as much like our world as possible. The overdetermination approach says that in such a world, the mental cause causes a physical event—namely, that the principle of causal closure of the physical no longer holds. I do not think we can accept this consequence: that a minimal counterfactual supposition like that can lead to a major change in the world. (45)

This is puzzling. Consider two worlds: in W1, *M* occurs and *N* does not but *M* is realized by another brain state *N** and both *M* and *M** are causes of *E*; in W2, *M* occurs and causes *E* but *E* has no physical cause. Kim seems to be saying that W2 is more similar to the actual world than W1 But why think this? In W1, supervenience and the completeness of physics both hold, while in W2 they both fail In fact, it seems obvious that we accept 'If *M* had occurred but had not been realized by *N*, it would have been realized by another neural state' and reject the corresponding counterfactual with the consequent ' supervenience and the completeness of physics would have failed'

---

[8] In previous writings, Kim formulated an "exclusion principle" according to which no event has two *independent* causal explanations and which he uses to argue against overdetermination. The exclusion principle figures a little less prominently in these lectures, perhaps because it does not seem applicable to mental-physical causation in NRP since, on it, mental and physical events are not metaphysically independent.

It becomes clear why Kim thinks that dependent overdetermination is ruled out when he describes what he thinks is the right way to describe the situation if NRP holds

> However, if we understand the difference between genuine, productive and generative causal processes, on the one hand, and noncausal regularities that are observed because they are parasitic on *real (sic)* causal processes, we are in a position to understand the picture recommended by (viii) (i e by NRP) In the case of the supposed M-M* causation the situation is rather like a series of shadows cast by a moving car The moving car represents a genuine causal process, but the series of shadows it casts, however regular and law-like it may be does not constitute a causal process (45)

Kim is thinking of causation as a relation in which the cause *generates* or *produces* the effect I am not sure what these metaphors come to, but they suggest that in some way, *E* grows out of *C* In any case, if we think of causation in this way then each of the reasons that Kim gives against overdetermination appears more convincing (1) If both *N* and *M* each produce *E*, then, even if they are metaphysically connected, they can do so independently of each other (2) If we have reason to believe that *N* produces *E*, then we have no need also to posit *M* as productive of *E* That hypothesis is dispensable (3) If *M* produces *E*, then, if we just take *N* away, *M* would still produce *E*, thereby violating the causal completeness of the physical If we reject overdetermination, then we shall have to say that *M* does not produce *M** (or anything else) but is like the shadow of the car So if we think of causation in terms of one event generating or producing another, then it looks like we may have good reason to hold onto (4) Finally, the metaphor of *C* producing *E* or *E* growing out of *C* suggests that it is impossible for *E* to grow out of any distinct *E* In the case of independent overdetermining causes, this would mean that the effect grows out of both, each contributing some aspects of the effect (for example, two bullet wounds) But in the case of NRP, it looks as though *M* contributes nothing more that *N* already contributes[9]

[9] As I mentioned, it is not clear whether Kim thinks that the nonexistence of overdetermining causes is a plausible claim based on an evidentiary principle 'Don't

Under the heading "free lunches," Kim discusses counterfactual accounts of causation as a way for NRP to satisfy (1-3) and motivate rejecting (4) The basic idea of the counterfactual account is that, on a certain understanding of counterfactuals, *C* causes *E* if *C* and *E* are distinct events (or propositions stating their occurrence) and *E* counterfactually depends on *C*, that is, $\neg C > \neg E$[10] On David Lewis's well-known account of counterfactuals, the completeness of physics is compatible with the overdetermination of counterfactual dependence So why is this not a way in which NRP can both have and eat its cake? Kim gives two reasons One is that the counterfactual account of causation is inadequate as an account of causation The other is that Lewis's account of counterfactuals—which may solve some of problems he raises for the counterfactual account—involves "heavy metaphysical assumptions" (130, footnote 19) Since Kim does not say what these assumptions are, I shall address only the first kind of worry[11]

Kim's main objection is that the counterfactual account does not distinguish epiphenomenal relations from genuine causation Let S1 be the location of the car's shadow at t1 and S2 at t2, and C1 and C2 the corresponding locations of the car The counterfactuals $\neg S1 \neg S2$ and $\neg S1 \neg C2$ may strike us as true, though there is no causal relation between S1 and either S2 or C2 Further, it seems to Kim that the supervenience of *M* on *N* is compatible with epiphenomenalism with respect to *E* (indeed Kim thinks this is the case if *M* is distinct from *N*) and yet in this situation $\neg M > \neg E$ So the counterfactual account cannot distinguish epiphenomenalism from mental causation Kim also observes that $\neg E > \neg E^*$ may be true when *E* does not cause *E**, but both are effects of a common cause and that $\neg E > \neg C$ may be true when *C* causes *E* If these objections are correct, then the counterfactual account fails

I think these objections can be overcome by combining an improved version of Lewis's account of counterfactuals with a slightly

---

multiply causes beyond necessity' or a conceptual/necessary truth It looks like Kim's argument against NRP requires the latter, since it is hard to see why this instance of the evidentiary principle is more plausible than the conjunction of (1-3)

[10] David Lewis—*Philosophical Papers, Volume II* (New York Oxford, 1986)—proposes a condition that is necessary as well as sufficient *C* causes *E* if and only if there is a chain of counterfactual dependencies from *C* to *E* There are well-known problems with the necessity of the account due to pre-emptive causes Lewis has recently modified his account greatly to deal with them, see "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000) 182-97 But whether or not this is successful if the condition is sufficient, and it could be shown that it permits mental causation that is compatible with NRP, Kim's objections would be answered

[11] Kim may have in mind Lewis's view that possible worlds are concrete totalities, but issues about the nature of possible worlds are orthogonal to issues concerning the adequacy of counterfactual accounts of causation

more complicated account of causation On Lewis's account, $A>B$ is true if and only if there are worlds at which A and B are true and no world at which $A\&\neg B$ is true is more similar to the actual world than is some world at which $A\&B$ is true Two factors determine the ranking of worlds as more or less similar to the actual world (i) the size of the regions of match in occurrent fact, and (ii) the size of the regions of conformity to laws of the actual world According to the account, 'If I had struck the match at *t* it would have lit at *t'*' is true since the most similar world(s) to the actual world is one that is exactly like the actual world until a time shortly before *t* when a small violation of the actual laws leads to my striking the match The laws then evolve that world so to entail that the match lights at *t'*

It is important to keep in mind that in evaluating Lewis's counterfactual account of causation, one must employ counterfactuals characterized by Lewis's proposed similarity relation and not merely appeal to intuitions It may be that we are inclined to affirm the counterfactual 'If her mother had not had blue eyes, then neither would she,' even though we know that it is not the mother's blue eyes but rather a gene factor that is a cause of the daughter's blue eyes But this is irrelevant as an objection to the counterfactual account of causation, since Lewis is characterizing a special class of counterfactuals and, on his characterization, this counterfactual (plausibly) comes out false [12] Lewis argues that his counterfactuals (and so the account of causation in terms of these counterfactuals) can handle the problem of the direction of causation and the problem of effects of a common cause [13] In order to handle the problem concerning epiphenomenalism for higher-level events, we need to complicate Lewis's account Say that M causes E if $\neg M>\neg E$ and there is no lower-level event which preempts this relation which is not itself preempted by M[14] N preempts M with respect to E if $\neg N>\neg E$ and

---

[12] That is, Lewis thinks that the most similar world in which the mother does not have blue eyes is one in which she still has the gene factor for blue eyes but something goes wrong in the phenotypic expression of that factor in her and that factor is passed on to the daughter who then has blue eyes

[13] In fact, as Adam Elga and Douglas Husak have recently argued in dissertations (at Massachusetts Institute of Technology and Rutgers University, respectively), Lewis's account of counterfactuals requires revision having to do with statistical mechanics But, this is not relevant to my defense of the role of counterfactuals in mental causation since there are repairs that can be made in the Lewis account to avoid the problems

[14] We also need to include that the physical states that are the supervenience bases for M and E are connected by fundamental dynamical law A proposal along these lines is found in Ernest Lepore and myself, "Mind Matters," this JOURNAL, LXXXIV, 11 (November 1987) 630-42, and in Lepore and my "More on Making Mind Matter More" *Philosophical Topics* (1991) 157-91

$N\&\neg M>E$ The intuitive idea is that M causes E if E counterfactually depends on N but does not counterfactually depend on M as long as N occurs This condition clearly makes the shadow epiphenomenal with regard to the car's position and the future shadow Both the mental epiphenomenalist and the interactionist dualist hold that mental events are (at most) connected by bridge laws to neurophysiological states [15] It follows that $N\&\neg M$ is metaphysically possible Suppose that $\neg M>\neg E$ so M is a putative cause of E The epiphenomenalist holds that it is not a genuine cause since N preempts M, while the interactionist dualist holds that M is not preempted and is a genuine cause The nonreductive physicalist holds that there are mental events M that are putative causes that are not preempted by events that they themselves do not preempt The existence of such mental events is compatible with nonreductive physicalism

I do not expect Kim to be persuaded by my counterfactual defense of causation He is likely to consider it "causation lite" when compared to causation as production or generation But the generation/production conception of causation fits ill with contemporary physics Bertrand Russell[16] famously said that causation so understood "is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously is supposed to do no harm" (*ibid*, p 1) What Russell had in mind is that the fundamental laws and facts of physics do not mention causality at all If one is a physicalist, then one holds that whatever causal relations there are supervene on the fundamental laws and facts of physics But the fundamental laws (for example, Schrodinger's law) relate the totality of the physical state at one instant to the totality at later instants [17] The laws do not single out parts of states at different times as being causally related If S' is the microphysical state in a region R at time t' and t' is a time prior to t, then nothing less than the state S of the region R^* that fills the backward light cone of R can be said to produce S' We cannot say that one event (one part of the physical state) produces another part since the laws do not connect parts in this way If we want an account of causation which selects among the parts of the state at t the parts that are responsible for a subsequent event, we have to look elsewhere One approach is to ask which events, given the laws, *make a*

---

[15] Such bridge laws do not supervene on the physical laws and states

[16] "On the Notion of Cause," *Proceedings of the Aristotelian Society*, XIII (1912)

[17] Or, a little more correctly, a region of the state at time t is related by law to the states in its backward light This point, which is often neglected by philosophers, is forcefully made in Noa Latham, "Singular Causal Statements and Strict Deterministic Laws," *Pacific Philosophical Quarterly*, LXVIII (1987) 29-43, and Hartry Field, "Causation in a Physical World" (manuscript)

difference to whether or not the effect occurs or how much of a difference it makes to what occurs This is the idea that counterfactual accounts attempt to capture I am not confident that Lewis's particular counterfactual account (with my additions) quite captures this idea But I do think that there must be some account on which causal relations, to the extent they are real, are determined by the fundamental laws and the noncausal fundamental facts On the other hand, the production conception of causation seems to require causal relations that fail to supervene on the fundamental laws (as they are conceived in contemporary physics) and facts That is, it seems to allow the existence of two universes that match perfectly in fundamental laws (of the sort posited by contemporary physics) and facts and yet differ in which events "produce" which others This indeed is a heavy metaphysical assumption

The problem of mental causation seems insoluble to Kim because he thinks of causation in terms of metaphors of generation and production This itself is a good reason to follow Russell's suggestion and dispense with this conception of causation I suggest that the difference making conception of causation suitably spelled out in terms of counterfactuals may be all we can get and all we need from causation [18] In this case, a free lunch may be preferable to indigestible metaphysics

BARRY LOEWER

Rutgers University

---

[18] There are still other problems involving mental causation The nonsupervenience of intentional mental states on brain states raises worries about whether the relevant counterfactuals that ground difference making causation are forthcoming Also, a satisfactory account of mental causation needs to spell out the mechanisms by which mental states are causally related to each other and to physical states Kim mentions these problems, but the book is almost entirely devoted to the problem presented by the incompatibility of (1-4)

Semantics, Tense, and Time PETER LUDLOW Cambridge Massachusetts Institute of Technology Press, 1999 xxi + 252 p Cloth $45 00

This is not your typical book about the A-theory/B-theory controversy in metaphysics Peter Ludlow attempts something that few philosophers have tried in the last thirty years he actually argues from linguistic premises for metaphysical conclusions The relevant linguistic premises have to do with the nature of language, a general theory of semantics, the proper analysis of tense, and various technical theses involving the treatment of temporal indexicals and temporal anaphora (among other things) The metaphysical conclusions that Ludlow argues for from these linguistic premises are some of the main claims normally associated with the A-theory in the philosophy of time, namely, (i) that tense is a genuine feature of the world (which means, roughly, that there are monadic, temporal properties—like pastness, presentness, futurity, being two days past, being three days future, and so on (sometimes called "A-properties")—and that the instantiation of these properties does not somehow reduce to the instantiation of two-place, temporal relations like earlier than, simultaneous with, later than, and so on), (ii) that temporal becoming (roughly, successively coming to possess different A-properties) is intrinsic to all events, and (iii) that only the present is real

The overall plan of the book is as follows First, Ludlow spends four chapters defending a cluster of related claims about language and semantics in general and, in particular, the semantics for temporal indexicals and temporal anaphora (He says that none of this material is original—he attributes most of it to Donald Davidson, Noam Chomsky, Gareth Evans, and James Higginbotham—but it seems to me that a fair portion of what goes into this part of the book (including, especially, Ludlow's way of combining various ideas) is indeed both original and of genuine significance) Then he devotes a chapter to developing a B-theory semantics, and another chapter to discussing what he takes to be certain insurmountable problems with the B-theory semantics Next, he devotes a chapter to developing an A-theory semantics, and another chapter to discussing what he takes to be the main problems facing those A-theory semantics, together with his reasons for thinking that those problems can be overcome The penultimate chapter of the book is devoted to arguing that considerations from "cognitive psychology generally" and "the theory of concept/language acquisition specifically" favor the A-theory over the B-theory And the final chapter contains remarks by Ludlow on what he takes to be the philosophical and linguistic consequences of

0022-362X/01/9806/325–29

his investigations There are also two "philosophical appendices" and five "technical appendices " (Taken together, the appendices constitute about one-fourth of the entire book)

Semantics, Tense, and Time is an impressive and ambitious book Ludlow touches on a huge number of different arguments, almost all of which raise interesting and important issues (In fact, some will find the book to be almost too ambitious, and will wish that Ludlow had focused on a smaller number of topics and arguments, in order to be able to consider in greater detail those arguments and the most likely rejoinders to them) There is no way I could do justice to even half a dozen of the most important arguments in the book, so in the remainder of this review, I shall limit my discussion to three main worries concerning Ludlow's treatment of some of his central topics

My first worry has to do with the relation between the A-theory, as formulated by Ludlow, and the claim—sometimes called *presentism*—that only present objects exist Ludlow formulates the A-theory in such a way that it incorporates presentism But in my mind, this is both unnecessary and problematic

Building presentism into the A-theory is unnecessary because, although presentism entails the A-theory, the opposite is not true That is, the essential claim of the A-theory—that tense is a genuine feature of the world—does not entail presentism For one could believe that tense is a genuine feature of the world—that various events, for example, have the property of being past, while others have the property of being present and still others have the property of being future—without thereby being committed to the claim that only present things exist In other words, it is quite possible to be an A-theorist but not a presentist (In fact, I suspect that the majority of A-theorists are not also presentists)

So there is no reason why we are forced to build presentism into the A-theory Moreover, doing so is problematic because of the many difficulties facing the presentist For example, Ludlow is forced to conclude that the A-theorist must avoid all reference to past and future things, including past and future times This turns out to be a major constraint on the different possible A-theory semantics that Ludlow considers, and it leads to all manner of difficulties for the A-theorist (For the record, I think that there are some important ways around this problem which are available to the presentist but not considered by Ludlow, including the construction of past and future times out of present resources (such as maximal propositions) [1])

---

[1] For more on this line, see Arthur Prior and Kit Fine, *World, Times and Selves* (London Duckworth, 1977), and Edward Zalta, "On the Structural Similarities between Worlds and Times," *Philosophical Studies* LI (1987) 213-39

My second main worry has to do with Ludlow's treatment of what he considers to be the biggest problem facing the B-theory, namely, the problem of temporal indexicals As Ludlow spells it out, the problem begins with an alarming episode from his personal life It seems that Ludlow is sitting at his desk one day (March 12, to be specific), unaware of the exact date, when he has the thought that his fifth anniversary is coming up sometime soon, and so calmly utters a token of this sentence

(1) My fifth anniversary is (this) March 12

A short time later, he happens to glance at his calendar, whereupon he goes into a panic, and frantically utters a token of the following sentence

(2) My fifth anniversary is today

Suppose we call these two utterances "U1" and "U2," respectively Then it should be clear that, intuitively, there is a huge difference between the meanings of U1 and U2

The A-theorist can explain our intuition that U1 and U2 have different meanings by invoking the putative A-property of presentness That is, he can say that U2 asserts that Ludlow's fifth anniversary has the property of being present, while U1 merely asserts that Ludlow's fifth anniversary and March 12 are simultaneous with each other And this difference in what the two utterances assert can be attributed to a difference in the meanings of U1 and U2

The B-theorist, on the other hand, is limited to just B-theory resources, so she cannot explain our intuition that U1 and U2 have different meanings in this way And, in fact, if we limit ourselves to considering just B-theory concepts, we seem forced to say that the two utterances say exactly the same thing

Ludlow considers various possible ways for the B-theorist to respond to this problem Each of the ways he considers is a semantical solution that involves trying to give "tenseless" truth conditions for sentences like U2 And each of the proposed solutions either fails miserably, or else leads to further difficulties, or both Thus, Ludlow concludes, consideration of the problem of temporal indexicals shows that the B-theory should be rejected

This appears at first glance to be a good argument against the B-theory, but I think that the argument turns out to be a little bit too good The problem of temporal indexicals is indeed a serious problem for the B-theorist [2] But, fortunately for the B-theorist, it is not a problem that is peculiar to the B-theory For as John Perry[3] has shown, the problem of temporal indexicals is a special case of a more general problem involving temporal, spatial, and personal indexicals alike

---

[2] See Prior, "Thank Goodness That's Over," in his *Papers in Logic and Ethics* (London Duckworth, 1976), pp 78–84, and Richard Gale, *The Language of Time* (London Routledge & Kegan Paul, 1968)

To see this, consider a variation on Ludlow's example Suppose that Ludlow is cycling in Vermont, and has promised to call his wife from the town of Ludlow immediately upon arriving there Late one morning, he arrives at a picturesque town and stops for lunch As he sits in the town square munching on a sandwich, his thoughts turn to his wife and home, and he calmly utters a token of this sentence

(3) I promised to call my wife as soon as I arrived in Ludlow

A short time later, a passerby happens to remark casually as he walks past, "Welcome to Ludlow," whereupon Ludlow (the man, not the whole town) goes into a panic, and frantically utters a token of the following sentence

(4) I promised to call my wife as soon as I arrived here

Now, it is clear that these last two utterances—call them U3 and U4, respectively—do not at first glance appear to have the same meaning But it is not at all clear that this should count as evidence for the claim that there are special, monadic, spatial properties like hereness

Or consider this example Suppose that Ludlow is suffering from amnesia, and cannot remember who he is or what he has done While puzzling over those topics, he is reading with great fascination from a book called *Semantics, Tense, and Time* He glances at the title page and calmly utters a token of this sentence

(5) Ludlow is the author of *Semantics, Tense, and Time*

A short time later, something in the book jogs his memory, and it all comes back to him, whereupon he excitedly utters a token of the following sentence

(6) I am the author of *Semantics, Tense, and Time*

I think that what Ludlow's example and these other examples all show is that there is something essential about indexicals, whether they be temporal, spatial, or personal indexicals But I do not think there is any reason to conclude that any such example shows that there must be ineliminable, monadic properties of the relevant kind Moreover, I think that, if Ludlow's argument from temporal indexicals succeeds, then it shows too much For there will be an analogous argument from spatial indexicals, whose conclusion is that there are ineliminable, monadic, spatial properties, and there will also be an analogous argument from personal indexicals, whose conclusion is that there are ineliminable, monadic, personal properties Since these are wildly implausible conclusions about the metaphysics of places and the metaphysics of people, we should resist the spatial and personal analogues of the argument from temporal indexicals And this suggests that, whether or not we have independent reasons for accepting the A-theory (such as, for example, the fact that it strikes many of us as obviously true), we should not endorse the argument from temporal indexicals for that theory

---

[3] "The Problem of the Essential Indexical," *Noûs*, XIII (1979) 3-21

The third of my main worries is a general one that concerns Ludlow's overall approach to metaphysics Ludlow explicitly says that he sees his entire project as an attempt to show that one can argue from the structure of human language to the nature of reality My worry is this Even if we grant that aspects of a natural language (including its semantics) reflect the metaphysical commitments of that language's speakers, we have no reason to think that those commitments must be correct After all, it is perfectly possible to have a language built around some false metaphysical views

To be fair, this is a worry that applies to a large number of well-known arguments in philosophy, many of which are rightfully taken with great seriousness But I do think that, insofar as an argument for a metaphysical conclusion has purely linguistic premises (as opposed to a mixture of metaphysical and linguistic premises), we should be highly suspicious of that argument And I worry that many of Ludlow's main arguments may fall into this category

One final note, on a related point The title of Ludlow's book is well chosen It accurately reflects the contents of the book, which is more about semantics than it is about tense, and more about tense than it is about time There are some philosophical purists who will be put off by this approach, and who will consider the book to be for linguists rather than philosophers But even if this is true, it is also true that *Semantics, Tense, and Time* is a remarkably rich sourcebook for both philosophers of time and linguists The book contains numerous fascinating cases—mainly in the form of sentences from natural languages—that present all kinds of perplexing puzzles and provocative problems for various theories in both the philosophy of time and linguistics

NED MARKOSIAN

Western Washington University

NOTES AND NEWS

The Editors regret the death of the eminent American philosopher Willard Van Orman Quine on December 25, 2000, at the age of 92. A native of Akron, Ohio, Professor Quine enjoyed a long and distinguished career in philosophy, making contributions in logic, set theory, philosophy of language, epistemology, and other fields. After receiving his B.A. in 1930 from Oberlin College, he began his doctoral studies at Harvard University, completing his Ph.D. under the guidance of Alfred North Whitehead in 1932. Quine was first appointed a Junior Fellow for the Society of Fellows at Harvard from 1933-1936, hired as an Instructor in 1936, promoted to full professorship in 1948, and received the Edgar Pierce Professorship in 1956. He remained on the faculty until his retirement in 1978. His most influential essays include "Truth by Convention" (1936), "New Foundations for Mathematical Logic" (1937), "On What There Is" (1948), "Two Dogmas of Empiricism" (1951), "Ontological Relativity" (1969), and "Epistemology Naturalized" (1969). A short list of his distinguished published texts comprises *A System of Logistic* (1934), *From a Logical Point of View* (1953), *Word and Object* (1960), *The Ways of Paradox and Other Essays* (1966), *Ontological Relativity and Other Essays* (1969), *The Roots of Reference* (1974), *Theories and Things* (1981), *Pursuit of Truth* (1990), and *From Stimulus to Science* (1995). Quine wrote *The Time of My Life: An Autobiography* in 1985.

The significance of Quine's contribution to philosophy occasioned numerous awards and honors. He became a Fellow of the American Academy of Arts and Sciences in 1949, and served as President of both the Association of Symbolic Logic, 1953-1955, and the American Philosophical Association, 1957. Quine was a corresponding Fellow for the British Academy beginning in 1959, and a corresponding member of the Instituto Brasileiro de Filosophia beginning in1963. He became a Fellow of the National Academy of Arts and Sciences in 1977, the Institut de France in 1978, and the Norwegian Academy of Sciences in 1979. He received the Nicholas Murray Butler Gold medal from Columbia University in 1970, the F. Polacky Gold Medal in 1991, and the Charles University Gold Medal in 1993. He was recipient of the Rolf Schock Prize in 1993, and the Kyoto Prize in 1996.

Quine, a frequent contributor to this *Journal*, has been variously described as "the foremost representative of naturalism in the second half of the twentieth century", "probably the most important American philosopher" since World War II, "our most distinguished systematic philosopher", and "one of America's preeminent thinkers."

REVUE INTERNATIONALE DE PHILOSOPHIE

Editor: Michel MEYER

Université de Bruxelles, C P 188
50, avenue F.D Roosevelt - 1050 Brussels, Belgium

Each number is devoted to a particular movement, a particular philosopher, or a particular problem

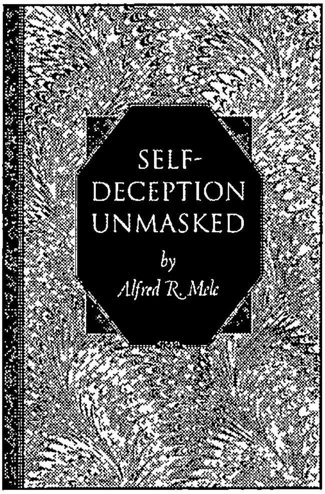
We publish 4 issues annually. Articles are written in English, French, German, or Italian

Our last issue was devoted to *Gadamer. Contemporary philosophers / Philosophes contemporains.*

Articles by Joel Weinsheimer, *Charity militant: Gadamer, Davidson, and Post-critical Hermeneutics* – Georgia Warnke, *Politics and literary criticism · a hermeneutic view* – Jean Greisch, *Le phénomène du jeu et les enjeux ontologiques de l'herméneutique* – Jean Grondin, *L'universalité de l'herméneutique et de la rhétorique. Ses sources dans le passage de Platon à Augustin dans* Vérité et méthode – Jaakko Hintikka, *Gadamer. squaring the hermeneutical circle* – Gianni Vattimo, *Histoire d'une virgule Gadamer et le sens de l'être* — Comptes rendus. Jacques Derrida, *Le Toucher, Jean-Luc Nancy* (Ann Van Sevenant) – Jean Bessière, *La littérature et sa rhétorique* (Philippe Daros)

Coming numbers: Philosophy of Economics – Bourdieu et la philosophie – Philosophy of Biology – L'Esthétique de Hegel – Sociologie des Sciences – Ryle – Searle – Le Corps – Putnam

New from Princeton



Self-Deception Unmasked
Alfred R. Mele

Self-deception raises complex questions about the nature of belief and the structure of the human mind In this book, Alfred Mele addresses four of the most critical of these questions What is it to deceive oneself? How do we deceive ourselves? Why do we deceive ourselves? Is self-deception really possible?

"This book presents a fine-grained, distinctive, and highly plausible account of self-deception in an engaging and concrete style "—Robert Audi

Princeton Monographs in Philosophy, Harry Frankfurt, Editor

Paper $15 95 ISBN 0 691-05745-1 Cloth $49 50 ISBN 0-691-05744-3 *Due January*

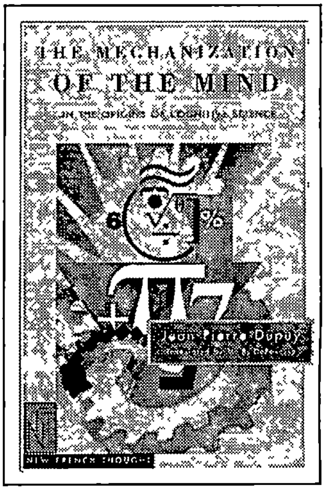
The Mechanization of the Mind
On the Origins of Cognitive Science
Jean-Pierre Dupuy
Translated by M B DeBevoise

A profound and beautifully written book, *The Mechanization of the Mind* brings back to life the intellectual brilliance and excitement that attended the birth of cognitive science more than fifty years ago

"Dupuy, in this brilliant translation by Malcolm DeBevoise [places] cybernetics within a centuries-long perspective of European philosophy which enriches our understanding of the successes and failures of present-day cognitive science "
—Michael Arbib

New French Thought, Thomas Pavel and Mark Lilla Editors

Cloth $29 95 ISBN 0-691-02574-6



Princeton University Press
800-777-4726 • WWW.PUP.PRINCETON.EDU